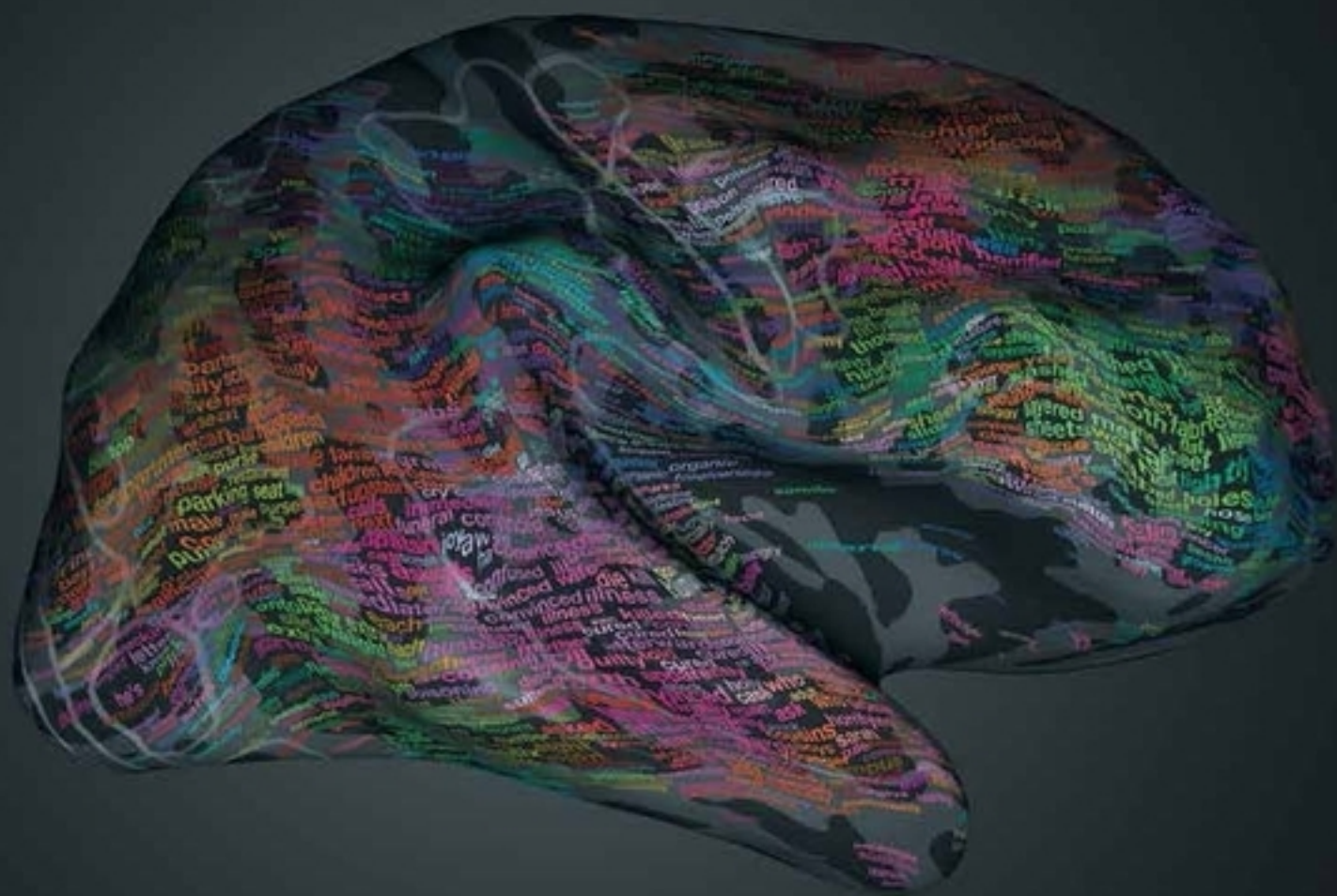


nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE



Where words make sense

A semantic atlas of the cerebral cortex

PAGE 453

NATURAL HAZARDS

MOVING MOUNTAINS

Listening for landslides after Nepal's killer quake

PAGE 428

INDUSTRIAL CHEMISTRY

TRIAL SEPARATIONS

Seven processes that could change the world

PAGE 435

PALAEONTOLOGY

SOME KIND OF VERTEBRATE

Demystification of the Illinois Tully monster

PAGES 447, 496 & 500

NATURE.COM/NATURE

28 April 2016 £10

Vol. 532, No. 7600



THIS WEEK

EDITORIALS

CANCER Moonshot project gets researchers on board **p.414**

WORLD VIEW Track small acts of aggression to fix science **p.415**



BOTANY Plant bleeding draws predators for protection **p.416**

Anticipating artificial intelligence

Concerns over AI are not simply fear-mongering. Progress in the field will affect society profoundly, and it is important to make sure that the changes benefit everyone.

In January, the Information Technology and Innovation Foundation in Washington DC gave its annual Luddite Award to “a loose coalition of scientists and luminaries who stirred fear and hysteria in 2015 by raising alarms that artificial intelligence (AI) could spell doom for humanity”.

The winners — if that is the correct word — included pioneering inventor Elon Musk and physicist Stephen Hawking.

In January last year, both signed an open letter that argued for research and regulatory and ethical frameworks to ensure that AI benefits humanity and to guarantee that “our AI systems must do what we want them to do”. Hardly “fear and hysteria”.

As AI converges with progress in robotics, cloud computing and precision manufacturing, tipping points will arise at which significant technological changes are likely to occur very quickly. Crucially, advances in robot vision and hearing, combined with AI, are allowing robots to better perceive their environments. This could lead to an explosion of intelligent robot applications — including those in which robots will work closely with humans.

Even academic debate on AI has tended to be polarized between sceptics and fanciful futurists. Yet there is an emerging middle-ground consensus that AI research is poised to have profound impacts on society. For those who remain sceptical that progress is imminent, bear in mind that Google, Toyota, Facebook, Microsoft and other companies are together pouring billions of dollars into AI and robotics research, which they see as the next frontier for profits (see page 422). Efforts to accelerate research must be accompanied by safeguards against the potential pitfalls of these powerful technologies.

Stuart Russell, a computer scientist at the University of California, Berkeley, who is well known for his deeply sceptical views on over-expectations of technological progress, is convinced that it is time to assess and mitigate potential risks. “Several technologies are reaching the level where they could be developed in potentially harmful directions,” says Russell, who was a driving force behind the open letter signed by Musk and Hawking.

So, what are the risks? Machines and robots that outperform humans across the board could self-improve beyond our control — and their interests might not align with ours. This extreme scenario, which cannot be discounted, is what captures most popular attention. But it is misleading to dismiss all concerns as worried about this.

There are more immediate risks, even with narrow aspects of AI that can already perform some tasks better than humans can. Few foresaw that the Internet and other technologies would open the way for mass, and often indiscriminate, surveillance by intelligence and law-enforcement agencies, threatening principles of privacy and the right to dissent. AI could make such surveillance more widespread and more powerful.

Then there are cybersecurity threats to smart cities, infrastructure and industries that become overdependent on AI — and the all too clear threat that drones and other autonomous offensive weapons

systems will allow machines to make lethal decisions alone.

The first wave of AI is already beginning to pervade our lives inconspicuously, from speech recognition and search engines to image classification. Self-driving cars and applications in health care are within sight, and subsequent waves could transform vast sectors of the economy, science and society. These could offer substantial benefits — but to whom?

Historically, automation in agriculture and industry has caused mass extinctions of jobs and led to profound societal changes — including rapid urbanization. But job losses have typically been more than compensated for by jobs created in the service and high-tech industries.

Many experts worry that AI and robots are now set to replace repetitive but skilled jobs that had been thought to be beyond machines, and it's not obvious where new jobs would come from. The spectre of permanent mass unemployment, and increased inequality that hits hardest along lines of class, race and gender, is perhaps all too real.

A society dependent on AI could yield broad benefits if increased wealth resulting from gains in productivity is shared. But currently, most such benefits are concentrated in companies and the capital of their shareholders — including the infamous 1%.

It is crucial that progress in technology is matched by solid, well-funded research to anticipate the scenarios it could bring about, and to study possible political and economic reforms that will allow those usurped by machinery to contribute to society. If that is a Luddite perspective, then so be it. ■

“As AI converges with progress in robotics, significant technological changes are likely to occur very quickly.”

On a downer

The United Nations has chosen to keep the war on drugs going — but it can't win.

Readers of the *Los Angeles Times* last week received some unexpected news about a major shift in the attitude of the United Nations towards the decriminalization of cannabis. According to the paper, the UN Office on Drugs and Crime (UNODC) was set to announce a more tolerant approach at a major meeting in New York City. But although the meeting was real, the policy shift was not. The announcement was a hoax, and pointedly timed for 20 April (‘4/20’), a day on which cannabis users celebrate and promote their choice. The scam even included a well-constructed fake press release

that quoted the (real) UNODC executive director Yury Fedotov as saying: “The science increasingly supports decriminalization and harm reduction over proscriptive, fear-based approaches.”

For those who advocate drug-law reform — a group that includes a sizeable number of scientists — the truth was a lot less encouraging. The comments that Fedotov made at last week’s UN General Assembly Special Session on Drugs (UNGASS) were certainly less quotable. In a tweet he noted: “#UNGASS outcome doc reaffirms joint responses to world drug problem based on agreed frameworks, #sharedresponsibility, intl cooperation”.

Despite hopes ahead of the meeting that nations would step back from the ‘war on drugs’ rhetoric that has defined international policy — and science — for decades, instead the UN blandly reformatted the existing status quo. Essentially, the message is still: ‘drugs are bad’.

This will disappoint the many readers of *Nature* who want to see a more evidence-based approach. And that disappointment is especially acute because hopes had been raised by a growing number of drug-policy experiments, such as legalization and decriminalization of cannabis in Uruguay and many US states.

If the overall message coming down from the highest levels remains the same, then so does the stance taken by those who fund research. Witness the struggles in the United States over cannabis studies: whereas some states permit citizens to openly smoke marijuana, researchers must wade through federal red tape to study it.

The harms that come from the current strategy of prevention through prohibition have been clearly demonstrated. Ahead of the meeting, researchers writing in *The Lancet* warned that the last UNGASS in 1998 made no distinction between drug use and drug misuse, leading to a focus on enforcement and a lack of focus on harm reduction (J. Csete *et al. Lancet* 387, 1427–1480; 2016).

This is not to say that drugs do not have risks or do not bring

damage. They can, and do, destroy lives and damage societies. Legalization brings its own problems — as places that have rushed to embrace commercial marijuana are finding out. The question is: what can be done to reduce harm and damage without creating more problems? And how can researchers find those answers? In other words, what would a reformed — and scientifically grounded — drug policy look like?

In January, the International Centre for Science in Drug Policy sent an open letter to the UN, signed by high-profile scientists from across the world, to ask the UNGASS to reconsider the metrics of drug use.

“Essentially, the message is still: ‘drugs are bad’.”

For too long, it said, countries have focused on a small number of metrics to judge the problem, including price, purity and levels of use in the general population. More-subtle indicators, such as treatment for drug-use disorders, drug-related murder and the proportion of prisoners

jailed for non-violent drug crimes, might be better metrics to measure, they suggested.

It will not surprise many people that there is a disconnect between drug policy and drug research. But discussions of drug policy, such as at UNGASS 2016, also seem to be increasingly out of step with the situation on the streets. The true picture of illegal drug use is, for obvious reasons, frequently opaque. But illegal drug use is clearly not in retreat. The billions spent, and the lives lost, in fighting the war on drugs have not brought the promised victories, and they are not likely to if the current course is maintained.

At the 1998 UNGASS, delegates pledged to deliver “significant and measurable” reductions in demand for drugs by 2008. That meeting even used the slogan: “A drug-free world, we can do it”. The deadline has slipped, but the intention seems to remain the same. Who are they kidding? ■

Biden time

The US vice-president’s cancer project is winning hearts and minds.

For many of the 18,000 people who were in New Orleans last week for the annual meeting of the American Association for Cancer Research, the highlight came when US vice-president Joseph Biden took the stage. Biden heads the US National Cancer Moonshot Initiative, which aims to double the pace of cancer research. He has consulted with hundreds of cancer researchers during his ‘listening tour’ to lay groundwork for the programme.

Biden seems to have been paying attention. He ran through a list of familiar obstacles posed by what he called “cancer politics” — the difficulties in conducting interdisciplinary research and sharing data, and the lack of incentives to reproduce published results, among others (see page 424). But it was when he made a joke about how long it takes to get a federal grant — “It’s like asking Derek Jeter to take several years off to sell bonds to build Yankee Stadium,” he said, referring to a famous baseball player — that it really hit home. The audience laughed and clapped; a few even gasped in surprise. The realization struck: the vice-president was clued up.

Biden made it clear that he was not the only one who was listening. At a recent nuclear-security summit with heads of state gathered round, US President Barack Obama began by noting that many of them had asked about Biden’s cancer initiative. Several countries, Biden said, then joined with the United States in a memorandum of understanding about how they could work together to fight cancer.

Are they right to be so enthusiastic? Certainly the flaws in Biden’s plan — not least the name — should not distract from its potential.

His National Cancer Moonshot Initiative could yet receive US\$1 billion in funding: not enough to ‘cure’ cancer, obviously, but perhaps enough to make significant changes in how cancer research is done if scientists help to target the money properly. And yes, the implications could yet spread beyond US borders — particularly if international researchers weigh in with their thoughts about how best to accelerate the pace.

The US National Cancer Institute has made it clear that it wants to hear recommendations from the community, and has a website dedicated to stimulating participation (see go.nature.com/cc5crk). This participation need not be restricted to US researchers: international scientists and clinicians should submit recommendations, too.

And, if the US project is as well received elsewhere as Biden claims, then scientists in those nations should look for ways to band together and marry their unique resources. Some countries have meticulous databases of health outcomes; others may have unique computing power or long-running longitudinal studies. And researchers in all countries face similar challenges of data sharing, reproducibility and interdisciplinary research.

These topics are also not cancer-specific: researchers in other fields have much to offer — and to gain. Biden said that after Obama’s State of the Union address, in which he appointed Biden head of the moonshot initiative, one of the first people to contact him was the US energy secretary Ernest Moniz. The Department of Energy has supercomputing power that could aid cancer researchers, the secretary said. Researchers from other fields can bring fresh perspectives to and reap the rewards of a coherent cancer-research strategy.

➔ **NATURE.COM**
To comment online,
click on Editorials at:
go.nature.com/xhunqv

In a US Congress that is paralysed by partisan bickering, the fight against cancer should find common support from lawmakers. Researchers can come together and show them the way. ■

JOHN ABROMOWSKI



Speak up about subtle sexism in science

Female scientists face everyday, often-unintentional microaggression in the workplace, and it won't stop unless we talk about it, says Tricia Serio.

Of all the questions I have been asked in my scientific career, perhaps the most troubling came from a former department head when I told him I was expecting my second child. "Was it planned?" he asked.

I had not yet secured tenure and took his remark to suggest that I was not committed to my career.

While I inwardly seethed at his assumption, I did not challenge it. Instead, like many women, I manoeuvre around such awkward and frequently offensive situations. In fact, at a women-in-science event at which I spoke, the organizer began by sharing strategies to change the subject when faced with inappropriate comments. But why should we? When such techniques are recommended as a form of professional development, enough is enough.

The problem of sexual harassment in science has been discussed in these pages and elsewhere, but less attention is paid to more indirect, subtle or unintentional comments. I think that this behaviour, sometimes known as microaggression, poses the greatest threat to diversity in science. Don't underestimate the sting and shock that these comments can cause: they make it quickly and painfully clear to women that, whereas we take situations at face value, others overlay our gender as a relevant consideration.

In my experience, these comments are infrequently discussed, and that's a missed opportunity. To improve the climate for women in science, it's time to share our stories of microaggression. Here are two more of my own.

On attending the first meeting of my linear-algebra class in college, the professor called my name from the roster (I was the only female student) and asked, "Why are you here?" And when I requested to meet a visiting professor with whom I was interested in discussing post-doctoral research opportunities, a professor and member of my PhD dissertation committee asked me, "Why? Jeff [my significant other at the time] is doing a postdoc in another city." At the time, I perceived these comments to mean, respectively: "You don't belong in this class" and "Your career is not a priority".

Microaggression arises in any situation in which there is a substantial demographic skew, so this problem is probably not specific to science, or gender. Nevertheless, my own anecdotes, and those from colleagues, suggest that they are prevalent and have an impact. Every woman, but not one man, whom I asked had a story to tell, but none had ever told it.

Unconscious gender bias is well documented in academic science. Women are entering the training pipeline in increasing numbers, but they exit more frequently than do men, leading to their under-representation in grants awarded and in academic positions.

Could microaggressions be driving women

from science? Inexcusably, I don't think we know.

Institutions have formal complaint mechanisms for people who have been subject to illegal gender discrimination and harassment. But microaggression may have the potential to cause more widespread harm. And because it doesn't seem to be actionable, examples often go unreported.

Although it is difficult to identify an innocuous reason why my former department chair felt it would be appropriate to comment on my family-planning decisions, I've come to believe that many microaggressions are voiced without an understanding of their impact on women. Regardless of their intended meaning, once spoken, they affect us. But our own silence also contributes to the problem. By not challenging and

discussing these comments, we miss the opportunity to educate the person making them, to decrease the chances of it happening again and to minimize their impact on us.

I know from experience that simply speaking out can make a difference on both sides. A few years ago, I purchased a set of soccer referee cards to use as a joke in conversations with a group of male professors. (They are friends but could do with some clarity on how their comments are perceived.)

The first time I issued a yellow card for a comment that questioned my maths skills, we all had a good laugh. But now they will pause and ask, "Was that a yellow or a red card?" and then explain what they actually meant (the questioning of my maths skills was apparently related to

my training as a biologist rather than my gender).

Heightening awareness of these communication gaps is the first step to diminishing their frequency and effects, and I propose a small and imperfect way to do just that. I invite those who have been subject to comments that they perceive as inappropriate, and those who have had their comments perceived incorrectly as inappropriate, to share those experiences anonymously on a website that I have created (www.speakyourstory.net). I will share these stories at regular intervals.

The purpose of this invitation is not to identify individual offenders. Rather, I hope to shine a light on the perception gap that I suspect leads to many microaggressions (and their subsequent impact), and to begin to quantify its scope by field, type of institution and location. My goal is to narrow or, ideally, to eliminate this gap. Let's inspire change by moving from unspoken anecdotes to awareness. Speak your story to pave the way. ■

Tricia Serio is professor and head of molecular and cellular biology at the University of Arizona in Tucson, and a public-voices fellow with the OpEd Project (www.theopedproject.org).
e-mail: tserio@email.arizona.edu

TO IMPROVE
**THE CLIMATE
FOR WOMEN**
IN SCIENCE,
IT'S TIME TO SHARE
OUR STORIES
OF MICROAGGRESSION.

➔ **NATURE.COM**
Discuss this article
online at:
go.nature.com/dkgf82

RESEARCH HIGHLIGHTS

Selections from the
scientific literature

AGROECOLOGY

Feed the world and keep the trees

A worldwide switch to vegetarian diets could allow the planet's estimated 2050 population of 9.7 billion to feed themselves without cutting down any more forests.

Karl-Heinz Erb and his colleagues at the University of Klagenfurt in Vienna created a model of the global agricultural system that forecasts the next 34 years, based on predictions of crop output per hectare, cropland expansion, efficiency of raising livestock, changes in the human diet and other variables. The team reports that given greatly increased crop yields and grazing intensity, global diets could stay much as they are without deforestation. A switch to a vegan or vegetarian diet could, however, allow sufficient expansion of even organically grown crops into former grazing land, without the need to boost yields.

Increased trade between areas of high production and high food demand will be needed to make any of these scenarios feasible.

Nature Commun. 7, 11382 (2016)

ASTRONOMY

Dwarf dark galaxy leaves smudge

Astronomers have found an elusive type of miniature galaxy.

Dwarf galaxies formed out of dark matter in the early Universe, but only a small number have been detected. Yashar Hezaveh of Stanford University in California and his colleagues studied images taken by the high-resolution Atacama Large Millimeter/

submillimeter Array (ALMA) in Chile. They found a galaxy that was acting like a lens, gravitationally bending light from another, more distant galaxy to form a ring of mirages in the images. The team spotted an additional 'smudge' on these mirages caused by an otherwise invisible dwarf galaxy orbiting the lensing galaxy.

The authors say that ALMA should be able to uncover more dark dwarf galaxies, which would bolster existing models of dark matter.

Astrophys. J. in the press; preprint at <http://arxiv.org/abs/1601.01388> (2016)



BOTANY

Plant 'bleeds' nectar from wounds

The sugary drops that form on the edges of wounds in a particular plant species have been identified as nectar, which attracts the enemies of leaf-eating pests.

Plants usually heal wounds quickly, but injuries to the bittersweet nightshade (*Solanum dulcamara*) do not close fully and produce a sugary secretion (pictured). Anke Steppuhn at the Free University of Berlin and her collaborators conducted greenhouse

experiments and found that the droplets attracted ants that defended the plant against two herbivorous pests: slugs and flea-beetle larvae.

Other plants produce nectar in specialized organs called nectaries, and the nightshade's organ-free way of making it could represent an evolutionary origin for these organs, the authors suggest.

Nature Plants <http://dx.doi.org/10.1038/nplants.2016.56> (2016)

NEUROSCIENCE

Brain may keep watch at night

Differences in brain activity between the left and right hemispheres may explain why people often sleep poorly in new environments.

Yuka Sasaki at Brown University in Providence, Rhode Island, and her colleagues imaged the brains of people sleeping in an unfamiliar setting, and measured slow-wave activity, a signal associated with non-rapid eye movement (NREM) sleep. During the first night, they found weaker activity

in the left hemisphere than in the right — an effect that disappeared on subsequent nights. The researchers also found that on the first night, sounds played to the right ear (which are processed by the left hemisphere) elicited greater brain activation and were more likely to wake the person up than sounds played to the left ear.

The authors speculate that lighter sleep in one hemisphere could have evolved out of a need for vigilance in new environments.

Curr. Biol. <http://dx.doi.org/10.1016/j.cub.2016.02.063> (2016)

TOBIAS LORTZING

DISEASE ECOLOGY

Map reveals global Zika risk

More than 2.17 billion people around the world live in habitats suitable for the mosquito-borne Zika virus, which is currently spreading in Central and South America.

Women infected with Zika virus when pregnant are at increased risk of giving birth to infants with microcephaly, which stunts brain development. Jane Messina at the University of Oxford, UK, and her colleagues used data on virus incubation, mosquito ranges and socio-economic variables associated with Zika outbreaks to produce a fine-scale global map that predicts areas with a high risk of virus spread (pictured, in red).

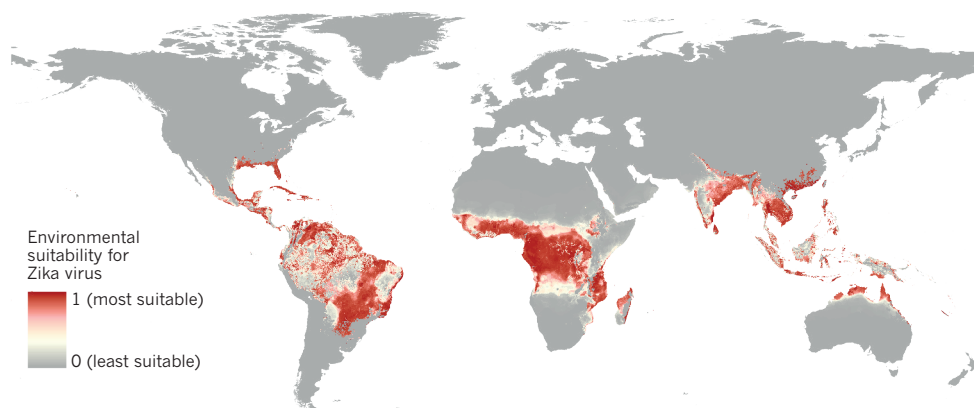
The map shows Florida and Texas to be ideal environments for Zika transmission, whereas southern Argentina, Uruguay and Chile, for example, are less likely to experience outbreaks. More than 5 million births in the Americas alone could be at risk from Zika infection over the next year, the authors say. *eLife* <http://doi.org/bfmg> (2016)

PHYSICS

Cold coffee beans grind smaller

Roasted coffee beans that are ground when cold give smaller particles than those ground at room temperature, which could affect the drink's flavour.

Christopher Hendon at the Massachusetts Institute of



Technology in Cambridge and his colleagues ground coffee beans (grinding burr pictured) and measured particle size, comparing beans from four countries and also those from a single source that were kept at four different temperatures before grinding. They found that particle-size distribution was similar across beans from different parts of the world. But particles became smaller and more uniform in size as the temperature dropped, with the largest change occurring between room temperature and -19°C .

Smaller and more uniformly sized coffee particles could release more flavour during brewing and might allow more coffee to be brewed from the same amount of grounds, the authors suggest.

Sci. Rep. 6, 24483 (2016)

GENETICS

'Welllderly' secrets revealed

A key to healthy ageing could be genetic protection against cognitive decline.

Eric Topol and Ali Torkamani of the Scripps Translational Science Institute in La Jolla, California, and their colleagues sequenced and analysed the genomes of more than 500 disease-free people over the age of 80. They found that these 'welllderly' individuals had a lower genetic risk of Alzheimer's disease and heart disease than a group of adults representing the general population, but about the same genetic risk of diabetes and cancer.

The authors suggest that welllderly people are genetically distinct from people who reach extreme old age by surviving significant health challenges. However, they add that the study was small and needs replication. Those in the welllderly group tended to be highly educated — a factor known to be linked to long lifespans, perhaps because of healthier lifestyles and diets.

Cell <http://dx.doi.org/10.1016/j.cell.2016.03.022> (2016)

METABOLISM

New hormone regulates glucose

Researchers have discovered a hormone that modulates the rapid release of glucose and insulin into the bloodstream in between meals.

Atul Chopra at Baylor College of Medicine in Houston, Texas, and his colleagues found that levels of the protein hormone — which they called asprosin — peaked in the blood during fasting in humans and mice, and is made by white fat tissue. Giving asprosin to mice boosted their blood glucose and insulin levels. The hormone bound to liver cells, triggering glucose release.

Mice and humans with insulin resistance (pre-diabetes) showed elevated levels of asprosin. Treating insulin-resistant mice with an asprosin-blocking antibody lowered their blood insulin levels. Decreasing asprosin could be a way to treat type 2 diabetes, the authors say. *Cell* 165, 566–579 (2016)

NEUROSCIENCE

How old age limits adaptability

Altered activity in two brain regions could explain why older animals struggle to adapt to changes in their environment.

Previous research has shown that a circuit connecting the thalamus and striatum helps animals to adapt their previous learning to a change in conditions. Jesus Bertran-Gonzalez at the University of Queensland in Brisbane, Australia, and his colleagues found that old mice showed weaker connections from the thalamus to the striatum, and had altered electrical properties in certain striatal neurons, compared with young mice. The team trained mice to press one lever to receive a preferred food and another for a non-preferred food; when researchers reversed these associations, old mice, as well as young mice with damage to this circuit, had trouble adapting their actions to the new rules.

The authors suggest that defects in this pathway impair an ageing animal's ability to integrate new and existing information when deciding how to act.

Neuron <http://dx.doi.org/10.1016/j.neuron.2016.03.006> (2016)

NATURE.COM

For the latest research published by Nature visit:

www.nature.com/latestresearch



SPENCER WEBB/GRINDSCIENCE.COM

SEVEN DAYS

The news in brief

RESEARCH

Contamination risk

Several clinical trials at US National Institutes of Health (NIH) facilities have been suspended in response to safety violations at manufacturing facilities. Two NIH labs — one that manufactures immune cells for use in patients at the National Cancer Institute in Bethesda, Maryland; the other that makes brain-imaging molecules at the National Institute of Mental Health in Bethesda — were shut down last week after contamination risks had been found. Twenty-eight clinical studies that use materials from these labs are on hold and will not recruit new patients until the issues are resolved, the NIH said on 19 April. See go.nature.com/uzq5kq for more.

Image problems

An analysis of more than 20,600 biomedical papers published from 1995 to 2014 in 40 journals has found that 4% contain deliberately or accidentally duplicated images (E. M. Bik Preprint at bioRxiv <http://doi.org/bfnw>; 2016). The prevalence of inappropriate images — and hence of misrepresented experiments — ranged from about 12% in the *International Journal of Oncology* to 0.3% in the *Journal of Cell Biology*. The authors of the study, a preprint of which was posted on 20 April, have reported all affected papers to the respective journals, so far resulting in 62 corrections and 6 retractions. See go.nature.com/axjb6l for more.

PEOPLE

Pachauri out

Rajendra Pachauri, former chair of the Intergovernmental Panel on



JEWEL SAMAD/AFP/GETTY

Paris climate pact signed

Representatives from more than 175 nations have signed the landmark Paris climate agreement to limit the rise in global average temperature to between 1.5°C and 2°C above pre-industrial levels. French President François Hollande (pictured, centre) and 54 other heads of state attended the signing ceremony on 22 April at the United Nations headquarters in

New York City, four months after the deal was agreed in Paris. By signing the treaty, which governments must yet ratify, nations formally pledge to reduce their greenhouse-gas emissions. However, current pledges to the United Nations are unlikely to keep warming below 2°C unless countries update their commitments in the near future. See go.nature.com/rezlw9 for more.

Climate Change (IPCC), has stepped down as executive vice-chair of the Energy and Resources Institute in New Delhi. Pachauri, who resigned last year as head of the IPCC amid accusations of sexual harassment, told the media that he wanted to pursue other interests.

Physicist in need

Jailed Iranian physicist Omid Kokabee underwent kidney-cancer surgery in Tehran on 20 April. Kokabee, who has studied laser physics in Spain and the United States, was arrested in Iran in 2011 while visiting his family, and sentenced to 10 years in prison for alleged espionage. The young scientist and his supporters claim that he was

sentenced for refusing to cooperate with Iran's nuclear programme. Appeals to Iran's supreme leader by the American Physical Society and the American Association for the Advancement of Science to release Kokabee on humanitarian grounds have previously gone unheeded.

FUNDING

Excellence drive

The German government plans to continue a multi-billion-euro initiative set up in 2005 to strengthen the research performance of the country's top universities. Starting in 2017, universities will be able to apply for extra government support from a total pot of €533 million

(US\$600 million) per year, federal research minister Johanna Wanka said on 22 April. In line with recommendations from an independent review, up to 50 research hubs will receive €3 million to €10 million a year. And 8–11 'excellence universities', which must host at least 2 such clusters, are to receive €10 million to €15 million per year.

ENVIRONMENT

Diversity wanted

A global effort to assess the value of biodiversity in ecosystems needs a broader range of expertise among the scientists who contribute to its reports. The Intergovernmental Science-Policy Platform on

Biodiversity and Ecosystem Services (IPBES) is struggling to find enough experts who could help it to assess the economic and social benefits of nature, IPBES chair Robert Watson warned last week. Earlier this year, a lack of social scientists — and of funding — prompted the panel to postpone its planned social-science programme for 2016, including a report on the diverse ways in which people value biodiversity. See go.nature.com/pa3dod for more.

Nuclear leak

An ongoing nuclear-waste leak has escalated at the Hanford Nuclear Reservation site in Washington state. On 18 April, the Washington Department of Ecology confirmed that radioactive waste is seeping from the primary tank of a double-shell storage container into the space between the primary and secondary tanks. Efforts to remove overflow waste safely from the tank were put on hold while engineers assess the situation. Officials emphasize that there is no indication of waste leaking into the environment.

Waning wolves

Ecologists involved in the world's longest-running predatory-prey study are mourning the declining



number of wolves in Isle Royale National Park (pictured). Wolves and moose on the island in Lake Superior, Michigan, have been studied for 58 years. But after several years of decline, the population of wolves dropped last year to just two inbred animals, and in the absence of predation, the moose population continues to rise. Isle Royale needs more wolves if the predators are to be an “ecological force” there, the team reported on 18 April (see go.nature.com/zlmjss).

onwards, which had raised concerns that it would exclude UK scientists from political consultations and democratic debate. In a 19 April statement, UK science minister Jo Johnson said that grants provided by the UK research and education councils and national academies will be exempted from the contentious gagging order. See go.nature.com/gewosj for more.

EVENTS

G7 science agenda

Brain science, disaster resilience and cultivation of young scientists are areas that urgently require a concerted global effort, science academies worldwide have told international leaders ahead of the G7 meeting next month in Japan. The statements, delivered to Japanese Prime Minister Shinzo Abe on 19 April,

POLICY

Free speech

Many scientists in the United Kingdom are to be exempted from a rule that will require recipients of public funding to maintain silence about the policy implications of their work. An ‘anti-lobbying’ clause will be applied to public grants awarded from May

COMING UP

30 APRIL – 3 MAY

The US National Academy of Sciences holds its 153rd annual meeting in Washington DC. go.nature.com/5hebou

2 MAY

US Department of Energy officials are expected to report to Congress on whether the United States should stay in ITER, the international fusion experiment under construction in Caderache, France. go.nature.com/vqrpqx

represent the opinions of learned academies from 13 countries, including the G7 nations, and the African regional science academy. The toll that brain disease has on well-being and the economy; increasing damage from natural disasters; and the need for well-trained scientists who can engage the public are areas of universal concern, the academies say.

BUSINESS

Biotech probe

The US government has launched investigations into the troubled blood-analysis firm Theranos of Palo Alto, California. *The Wall Street Journal* reported on 18 April that Department of Justice federal prosecutors and the government's Securities and Exchange Commission are examining whether Theranos misled investors. The company claimed to have developed new technology to allow testing of minute quantities of blood, but has since been accused of running most of its analyses on conventional platforms.

➔ NATURE.COM

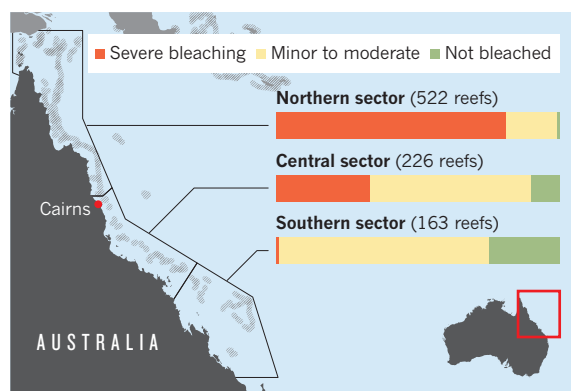
For daily news updates see: www.nature.com/news

TREND WATCH

Aerial and underwater surveys have revealed that 93% of Australia's Great Barrier Reef has been affected by bleaching, in the worst-ever observed event. Bleaching — in which heat-stressed corals expel their symbiotic algae — can kill corals when it is severe. The Australian Research Council Centre of Excellence for Coral Reef Studies in Townsville, Queensland, found that only 68 of 911 reefs have escaped bleaching, and hundreds have been severely affected. See go.nature.com/xamggr for more.

GREAT BARRIER REEF GROWS PALE

Bleaching has hit many reefs in the Great Barrier Reef. The northern parts are worst affected, with all corals bleached in some reefs.



NEWS IN FOCUS

TECH Excitement and concern as AI migrates to industry **p.422**

HEALTH Cancer ‘moonshots’ are lacking joined-up thinking **p.424**

PHYSICS Europe quietly plans a quantum revolution **p.426**



GENE EDITING How CRISPR upended Emmanuelle Charpentier’s life **p.428**

RUTH FREMSON/NYT/EYEVINE



Wheat in northern India could be threatened by an outbreak of fungal disease in Bangladesh.

AGRICULTURE

Devastating wheat fungus appears in Asia for first time

No one knows the origin of the Bangladesh outbreak, which scientists warn could spread.

BY EWEN CALLAWAY

Fields are ablaze in Bangladesh, as farmers struggle to contain Asia’s first outbreak of a fungal disease that periodically devastates crops in South America. Plant pathologists warn that wheat blast could spread to other parts of south and southeast Asia, and are hurrying to trace its origins.

“It’s important to know what the strain is,” says Sophien Kamoun, a biologist at the Sainsbury Laboratory in Norwich, UK, who has created a website, Open Wheat Blast (go.nature.com/bkczwf), to encourage researchers to share data.

Efforts are also under way to find wheat genes that confer resistance to the disease.

First detected in February and confirmed with genome sequencing by Kamoun’s lab this month, the wheat-blast outbreak has already caused the loss of more than 15,000 hectares of crops in Bangladesh. “It’s really an explosive, devastating disease,” says plant pathologist Barbara Valent of Kansas State University in Manhattan, Kansas. “It’s really critical that it be controlled in Bangladesh.”

After rice, wheat is the second most cultivated grain in Bangladesh, which has a population of 156 million people. More broadly,

inhabitants of south Asia grow 135 million tonnes of wheat each year.

Wheat blast is caused by the fungus *Magnaporthe oryzae*. Since 1985, when scientists discovered it in Brazil’s Paraná state, the disease has raced across South America.

The fungus is better known as a pathogen of rice. But unlike in rice, where *M. oryzae* attacks the leaves, the fungus strikes the heads of wheat, which are difficult for fungicides to reach. A 2009 outbreak in wheat cost Brazil one-third of that year’s crop. “There are regions in South America where they don’t grow wheat because of the disease,” Valent says. Wheat blast was ►

► spotted in Kentucky in 2011, but vigorous surveillance helped to stop it spreading in the United States.

In South America, the disease tends to take hold in hot and humid spells. Such conditions are present in Bangladesh, and the disease could migrate across south and southeast Asia, say plant pathologists. In particular, it could spread over the Indo-Gangetic Plain through Bangladesh, northern India and eastern Pakistan, warn scientists at the Bangladesh Agricultural Research Institute (BARI) in Nashipur.

Bangladeshi officials are burning government-owned wheat fields to contain the fungus, and telling farmers not to sow seeds from infected plots. The BARI hopes to identify wheat varieties that are more tolerant of the fungus and agricultural practices that can keep it at bay, such as crop rotation and seed treatment.

It is unknown how wheat blast got to Bangladesh. One possibility is that a wheat-infecting strain was brought in from South America, says Nick Talbot, a plant pathologist at the University of Exeter, UK. Another is that an *M. oryzae* strain that infects south Asian grasses somehow jumped to wheat, perhaps triggered by an environmental shift: that is what happened in Kentucky, when a rye-grass strain infected wheat.

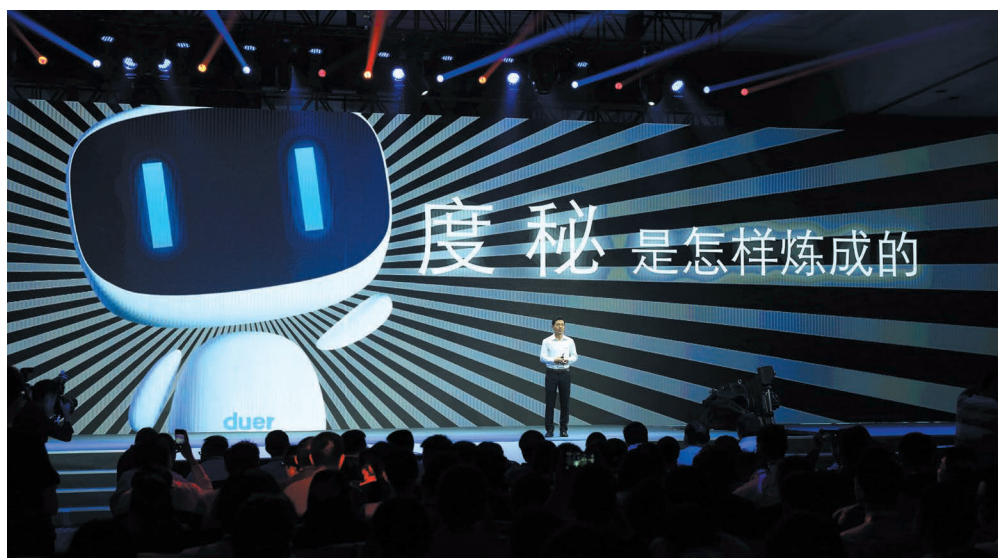
To tackle the question, this month Kamoun's lab sequenced a fungus sample from Bangladesh. The strain seems to be related to those that infect wheat in South America, says Kamoun, but data from other wheat-infecting strains and strains that plague other grasses are needed to pinpoint the outbreak's origins conclusively.

The Open Wheat Blast website might help. Kamoun has uploaded the Bangladeshi data, and Talbot has deposited *M. oryzae* sequences from wheat in Brazil. Talbot hopes that widely accessible genome data could help to combat the outbreak. Researchers could use them to screen seeds for infection or identify wild grasses that can transmit the fungus to wheat fields.

Rapid data sharing is becoming more common in health emergencies, such as the outbreak of Zika virus in the Americas. Kamoun and Talbot say that their field should follow suit. "The plant-pathology community has a responsibility to allow data to be used to combat diseases that are happening now, and not worry too much about whether they may or may not get a *Nature* paper out of it," says Talbot.

Last month, Valent's team reported the first gene variant known to confer wheat-blast resistance (C. D. Cruz *et al.* *Crop Sci.* <http://doi.org/bfk7>; 2016), and field trials of crops that bear the resistance gene variant have begun in South America. But plant pathologists say that finding one variant is not enough: wheat strains must be bred with multiple genes for resistance, to stop *M. oryzae* quickly overcoming their defences.

The work could help in the Asian crisis, says Talbot. "What I would hope for out of this sorry situation," he says, "is that there will be a bigger international effort to identify resistance genes." ■



Robin Li, head of China's web giant Baidu, unveils the firm's intelligent digital assistant, Duer.

ARTIFICIAL INTELLIGENCE

AI firms lure academics

Shift to industry sparks excitement — and some concern.

BY ELIZABETH GIBNEY

When Andrew Ng joined Google from Stanford University in 2011, he was among a trickle of artificial-intelligence (AI) experts in academia taking up roles in industry.

Five years later, demand for expertise in AI is booming — and a torrent of researchers is following Ng's lead. The laboratories of tech titans Google, Microsoft, Facebook, IBM and Baidu (China's web-services giant) are stuffed with ex-university scientists, drawn to private firms' superior computing resources and salaries. "Some people in academia blame me for starting part of this," says Ng, who in 2014 moved again to become chief scientist at Baidu, working at the company's research lab in California's Silicon Valley.

Many scientists say that the intense corporate interest is a boon to AI — bringing vast engineering resources to the field, demonstrating its real-world relevance and attracting eager students. But some are concerned about the more subtle impacts of the industrial migration, which leaves universities temporarily devoid of top talent, and could ultimately sway the field towards commercial endeavours at the expense of fundamental research.

Private firms are investing heavily in AI — and in particular in an AI technique called deep learning — because of its promise to glean understanding from huge amounts of data. Sophisticated AI systems might be able to create effective digital personal assistants, control self-driving cars, or take on a host of other tasks that are too complex for conventional programming. And corporate labs' resources allow progress that might not be possible in academic departments, says Geoffrey Hinton, a deep-learning pioneer at the University of Toronto in Canada who took up a post at Google in 2013. The fields of speech and image recognition, for instance, had been held up for years by a lack of data to use in training algorithms and a shortage of hardware, he says — bottlenecks that he was able to get past at Google.

"AI is so hot right now. There are so many opportunities and so few people to work on them," says Ng, who says he was attracted by Google's reams of data and computing power, and its ability to tackle real-world problems. Another temptation is that private firms can offer "astronomical" wages, says Tara Sinclair, chief economist at Indeed, a firm headquartered in Austin, Texas, that aggregates online job listings and has chronicled a rising demand for jobs in AI in Britain and the United States.

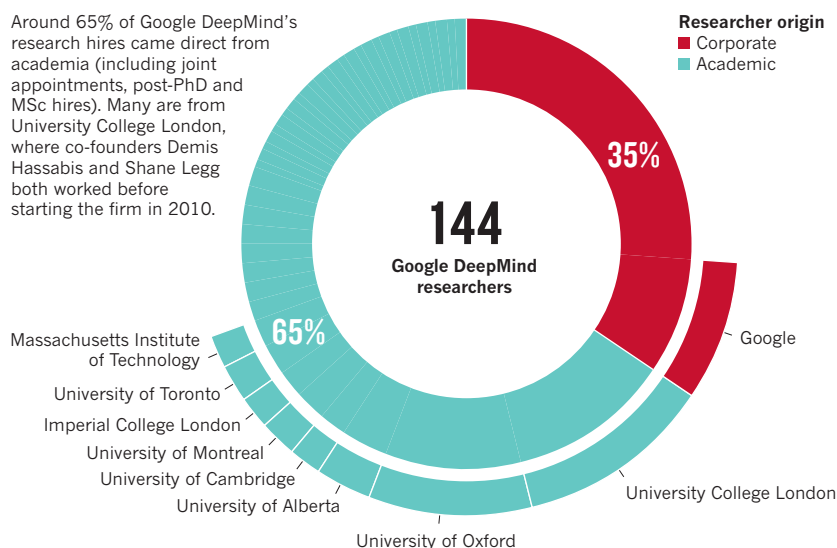
The excitement shows that AI is at a point at

CHINA/GETTY

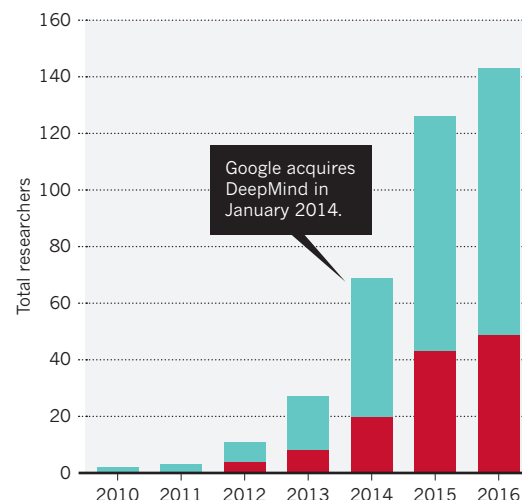
GOOGLE DEEPMIND'S TALENT GRAB

Google DeepMind, an artificial-intelligence firm in London, has embarked on a hiring spree since 2014. Its staff declined to discuss the migration of AI talent, but data gathered by Nature suggest that the firm's current roster includes at least 144 researchers — almost two-thirds of them drawn from universities.

Around 65% of Google DeepMind's research hires came direct from academia (including joint appointments, post-PhD and MSc hires). Many are from University College London, where co-founders Demis Hassabis and Shane Legg both worked before starting the firm in 2010.



Cumulative growth in Google DeepMind research staff.



Data collated by Nature from online sources including Scopus, LinkedIn, Google Scholar and personal webpages. 'Researchers' excludes most software engineers and developers, and all administrative or other staff. Researchers were identified by title (such as 'research scientist' or 'research engineer') or previous role. Not all institutions shown in breakdown.

which it can achieve real-world impact — and companies are the natural way to make this happen, says Pieter Abbeel, a specialist in AI and deep learning at the University of California, Berkeley. In the 1950s, a similar job migration occurred in semiconductor research, when many of the field's leading figures were poached to become heads of industrial research-and-development labs, says Robert Tijssen, a social scientist at Leiden University in the Netherlands. Moving academics bring expertise while extending their new-found corporate networks back to their former colleagues and students — making the scenario a "classic win-win situation," Tijssen says.

CORPORATE COLLABORATIONS

Herman Herman, director of the US National Robotics Engineering Center based at Carnegie Mellon University in Pittsburgh, Pennsylvania, subscribes to that view. In 2015, car-hailing app Uber, which was collaborating with the centre, hired almost 40 of his 150 researchers, mainly those working on self-driving cars. Reports at the time suggested that the centre was left in crisis, but Herman says this was overblown; the project was one of dozens across Carnegie Mellon's Robotics Institute, which has about 500 faculty members. The move was a chance to bring in new blood, and shortly afterwards, Uber donated US\$5.5 million to support student and faculty fellowships at the institute. Meanwhile, the publicity raised the profile of the centre's work, says Herman — and student applications are up.

The loss of expertise in academia concerns Yoshua Bengio, a computer scientist at the

University of Montreal in Canada, which has also seen a surge in graduate-student applications. If industry-hired faculty members do retain university roles — as Hinton has at the University of Toronto and Ng has at Stanford University in California — they are usually only minor, says Bengio. Losing faculty members reduces the number of students that can be trained, especially at PhD level, adds Abbeel.

Hinton predicts that the shortage in deep-learning experts will be temporary. "The magic of graduate research in universities is something to be protected, and Google recognizes that," he says. Google is currently funding more than 250 academic research projects and dozens of PhD fellowships.

In supplying industry with talent, universities are fulfilling their natural role, says Michael Wooldridge, a computer scientist at the Uni-

"There are so many opportunities and so few people to work on them."

versity of Oxford, UK. And with interest in AI generally booming, he struggles to see a situation in which academia is left deserted. The London-based firm Google DeepMind hired ten researchers from Oxford in 2014 — but Google gave the university a seven-figure financial contribution, and formed a research collaboration (see 'Google DeepMind's talent grab'). Many of the poached staff still hold active teaching positions at the university — giving students opportunities they might otherwise never have had.

Bengio is also concerned about the long-term impacts of corporate domination.

Industry researchers are more secretive, he says. Although scientists in some corporate labs (such as those at Google and Baidu) are still publishing papers and code openly — allowing others to build on their work — Bengio argues that corporate researchers still often avoid discussing their work ahead of publication because they are more likely than academics to have filed for patents. "That makes it more difficult to collaborate," he says.

Some industry insiders are concerned about transparency, too. In December 2015, SpaceX founder Elon Musk was among a group of Silicon Valley investors who launched a non-profit firm called OpenAI in San Francisco, California. With \$1 billion promised by its backers, it aims to develop AI for the public good, sharing its patents and collaborating freely with other institutions.

Although Google, Facebook and the like seem committed for the moment to tackling fundamental questions in AI, Bengio fears that this might not last. "Business tends to be pulled to short-term concerns. It's in the nature of the beast," he says. He cites telecommunications firms Bell Labs and AT&T as examples of companies that had strong research labs, but eventually lost their talent by putting too much emphasis on the short-term objective of making money for the company.

Hinton insists that basic research can thrive in industry. And because of the urgent need for AI research, some of today's expansion in basic research is inevitably taking place at corporate firms, he adds. But academia will still play a crucial part in AI research, he says. "It's the most likely place to get radical new ideas." ■



US vice-president Joe Biden speaks to university researchers about his cancer initiative.

US vice-president Joe Biden, said last week at the annual meeting of the American Association for Cancer Research (AACR) in New Orleans, Louisiana.

An advisory panel will release more-detailed plans for the government programme in June. Meanwhile, three privately funded immunotherapy research projects are gearing up: the \$250-million Parker Institute for Cancer Immunotherapy, funded by Sean Parker, co-founder of the music-file-sharing company Napster, and announced on 13 April; a \$125-million Immunotherapy Center at Johns Hopkins University in Baltimore, Maryland, unveiled in March; and the Cancer MoonShot 2020 Program, announced in January by biotechnology billionaire Patrick Soon-Shiong.

This sudden proliferation of cancer initiatives is reminiscent of the spate of brain-research projects launched in the past few years — some of which have floundered through poor leadership. Europe's Human Brain Project, for instance, almost ran aground after a series of top-down decisions alienated the neuroscience community. By contrast, the US BRAIN Initiative set priorities after consulting with neuroscientists, and awarded grants through a conventional peer-reviewed process, ensuring community acceptance.

Now cancer researchers are left wondering how their moonshots will proceed. At the AACR meeting, Biden said that he had met representatives of many cancer-funding projects. "Why is all of that being done separately?" he asked scientists in the audience, noting that progress is accelerated by collaboration.

The privately funded initiatives are more concerned with meeting their own goals — and satisfying their funders — than with coordinating efforts in the field. "I don't see my role as trying to answer this larger question about how does this all fit together," says Jeffrey Bluestone, chief executive of the Parker Institute. "I'm focused on how to make sure what we do is impactful for patients."

But Douglas Lowy, acting director of the US National Cancer Institute (NCI), which is coordinating the government moonshot, notes an overlap with the leadership of the various projects. Soon-Shiong, Bluestone and leaders of immunotherapy initiatives at Johns Hopkins and the University of Texas MD Anderson Cancer Center in Houston are on the

BIOMEDICAL RESEARCH

Cancer moonshots raise concerns

Scientists worry that US government and private funders are working at cross purposes.

BY ERIKA CHECK HAYDEN

The recent launch of multiple major US cancer initiatives has infused cash into immunotherapy, one of the most promising new methods of cancer treatment. But researchers warn that the money may be wasted without concrete plans to coordinate the programmes.

"There's a lack of overt leadership, and in the absence of a logical strategy we have a tendency to throw plates of spaghetti against

the wall and hope it sticks," says Ira Mellman, vice-president of cancer immunology at the biotechnology company Genentech in South San Francisco, California.

The broadest programme is the US government's National Cancer Moonshot, which hopes to receive US\$1 billion by next year for 8 areas of cancer research. Immunotherapy, which recalibrates the body's own immune defence against cancer, is among them. It "is poised to be a critical part of our nation's anticancer strategy", the project's leader,



**MORE
ONLINE**

TOP STORY



Why transgenic insects are still not ready for prime time go.nature.com/5ymmtr

MORE NEWS

- Plant protein behaves like a prion go.nature.com/lefzxi
- Evolution of Darwin's finches tracked at genetic level go.nature.com/5tybkl
- Genetic secrets of the healthy elderly unveiled go.nature.com/hddnqz

NATURE PODCAST



A language map of the brain, listening for landslides, and the Soviet internet that never was nature.com/nature/podcast

BIOMEDICAL RESEARCH

government initiative's advisory panel. And on 18 April, the Biden moonshot launched a website to solicit research ideas. The aim, Lowy says, is to ensure that research areas recommended by the advisory panel do not duplicate topics being covered by the private initiatives.

There is wide agreement on major questions regarding immunotherapy, however. For instance, researchers don't understand why the approach works in only 15–20% of patients. Combining immunotherapies, and studying what distinguishes patients who respond, could make treatments more effective. Pharmaceutical companies are already developing new drugs and testing therapies in combination. Philip Gotwals, executive director of oncology research at the Novartis Institutes for BioMedical Research in Cambridge, Massachusetts, estimates that industry has spent upwards of \$1 billion on the field.

But scientists see a lack of basic cancer immunology research, even in the new programmes. "Many of these initiatives are moving forward ideas that are already out there," says David Raulet, faculty director of the Immunotherapeutics and Vaccine Research Initiative at the University of California, Berkeley, which began in March.

Many researchers are looking to the Biden project to make a big investment in basic cancer immunology and to address broader barriers to research, such as data hoarding. Gotwals, for instance, notes that the results of industry-sponsored clinical trials now under way could help other companies to decide which approaches to test, but that results are typically not made public until 9–12 months after a trial ends. Companies are reluctant to share data before then, both to comply with regulatory requirements and to protect their intellectual property. "It's not trivial to figure out how to make that work," Gotwals says.

Biden seems to be hearing that message. At the AACR meeting, he said that data sharing often comes up when he speaks to scientists about the moonshot. Lowy says that the NCI is already planning to open a Genomic Data Commons in June to host detailed information on cancer patients. Sharing data collected in company-sponsored clinical trials is trickier because patients must give informed consent.

In the meantime, the government moonshot faces a major hurdle: its funding is at the mercy of legislators who may be loath to give US President Barack Obama a victory in his last year in office. "It will be very difficult for us to initiate all of the programmes that we're looking forward to the blue-ribbon panel recommending if there isn't funding," Lowy says. ■ [SEE EDITORIAL P.414](#)

Additional reporting by Heidi Ledford

Personalized tumour vaccines spark debate

Enthusiasm comes amid worries that the therapy may prove too complex to manufacture.

BY HEIDI LEDFORD

It is precision medicine taken to the extreme: cancer-fighting vaccines that are custom designed for patients according to the mutations in their individual tumours. With early clinical trials showing promise, that extreme could one day become commonplace — but only if drug developers can scale up and speed up the production of their tailored medicines.

The topic was front and centre at the American Association for Cancer Research (AACR) annual meeting in New Orleans, Louisiana, on 16–20 April. Researchers described early data from clinical trials that suggested that personalized vaccines can trigger immune responses against cancer cells. Investors seem optimistic that those results will translate into benefits for patients; over the past year, they have pumped cash into biotechnology start-up companies that are pursuing the approach.

But some researchers worry that the excitement is too much, too soon for an approach that still faces many technical challenges. "What I do really puzzle at is the level of what I would call irrational exuberance," says Drew Pardoll, a cancer immunologist at Johns Hopkins University in Baltimore, Maryland.

The concept of a vaccine to treat cancer has intrinsic appeal. Some tumour proteins are either mutated or expressed at different levels than in normal tissue. This suggests that the immune system could recognize these unusual proteins as foreign if it were alerted to their presence by a vaccine containing fragments of the mutant protein. The immune system's army of T cells could then seek out and destroy cancer cells that bear the protein.

Decades of research into cancer-treatment vaccines have thus far yielded disappointing clinical-trial results, but advances in recent years — including a suite of drugs that might amplify the effects of cancer vaccines — have rekindled hope for the field. And DNA sequencing of tumour genomes has revealed a staggering diversity of mutations that produce proteins that could serve as 'antigens' — molecules that alert the immune system to the tumour's presence.

Last year, researchers reported that they had triggered immune responses in three people with melanoma by administering a vaccine

tailored to their potential tumour antigens (B. M. Carreno *et al. Science* **348**, 803–808; 2015). The vaccines' effects on tumour growth are not yet clear, but by the end of 2015, several companies had announced their intention to enter the field. Gritstone Oncology, a start-up firm in Emeryville, California, raised US\$102 million to pursue the approach, and Neon Therapeutics of Cambridge, Massachusetts, raised \$55 million. A third company, Caperna, was spun out of the prominent biotechnology company Moderna Therapeutics, also in Cambridge.

VACCINE HURDLES

Academic groups are also moving quickly. At the AACR meeting, immunologist Robert Schreiber of Washington University in St. Louis, Missouri, described six ongoing studies at his institution, in cancers ranging from melanoma to pancreatic. And Catherine Wu of the Dana-Farber Cancer Institute in Boston, Massachusetts, presented

data showing signs of T-cell responses to a melanoma vaccine.

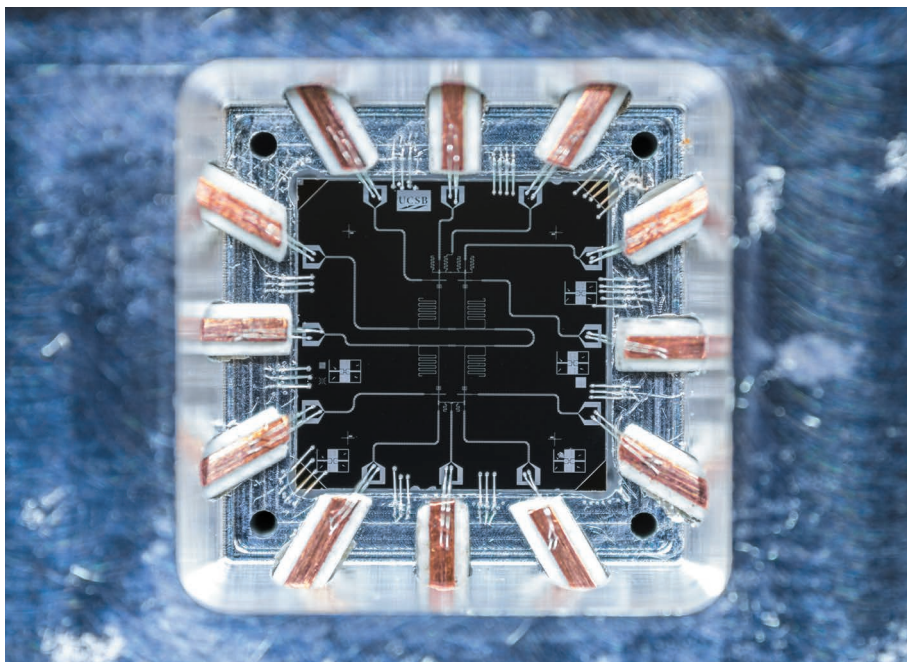
But it takes Wu's team about 12 weeks to generate a vaccine, and the Washington University team

needs about 8 weeks. That could limit the treatment to slow-growing cancers, says Wu.

There is also a reason that so many researchers choose melanoma for proof-of-principle trials. Melanoma tumours tend to harbour many mutations, which provide scientists with ample opportunity to select those that may serve as antigens. Some researchers worry that tumours with fewer mutations may not be so suitable for personalized vaccines.

Pardoll, meanwhile, is concerned that the field is shifting too quickly, leaving behind decades of research on antigens that might be shared across tumours — work that has not borne out in clinical trials thus far, but would be much simpler to manufacture and deploy on a large scale. "I will be the happiest person in the world to be proven wrong on these," he says of personalized vaccines. "But I think one has to nonetheless be cognizant of where the challenges are." ■

"What I do really puzzle at is the level of what I would call irrational exuberance."



A €1-billion (US\$1.1-billion) European flagship project could advance the state of quantum computing.

FUNDING

Billion-euro boost for quantum tech

Third European Union flagship project will be similar in size and ambition to graphene and human-brain initiatives.

BY ELIZABETH GIBNEY

The European Commission has quietly announced plans to launch a €1-billion (US\$1.1-billion) project to boost a raft of quantum technologies — from secure communication networks to ultra-precise gravity sensors and clocks.

The initiative, to launch in 2018, will be similar in size, timescale and ambition to the two existing European flagship projects, the decade-long Graphene Flagship and the Human Brain Project — although the exact format has yet to be decided, Nathalie Vandystadt, a commission spokesperson, told *Nature*. Funding will come from a mix of sources, including the commission, as well as other European and national funders, she added.

The commission is likely to have a “substantial role” in funding the flagship, says Tommaso Calarco, who leads the Integrated Quantum Science and Technology centre at the Universities of Ulm and Stuttgart in Germany. He co-authored a blueprint behind the initiative, which was published in March,

called the Quantum Manifesto. Countries around the world are investing in these technologies, says Calarco — without such an initiative, Europe risks becoming a second-tier player. “The time is really now or never.”

On 19 April, the commission formally announced its intention to support the initiative. Confusingly, the project is included under plans to launch a cloud-computing portal called the European Open Science Cloud, even though the remit of the quantum project will extend far beyond computing. (In the same announcement, the commission said that it would spend €2 billion on the cloud-computing initiative by 2020.)

QUANTUM BUZZ

High-profile US companies are already investing in quantum computing, and Chinese scientists are nearing the completion of a 2,000-kilometre-long quantum-communication link — the longest in the world — to send information securely between Beijing and Shanghai.

In Europe, the flagship is expected to fuel the

development of such technologies, which the commission calls part of a “second quantum revolution” (the first was the unearthing of the rules of the quantum realm, which led to the invention of tools such as lasers and transistors).

The initiative will include support for relatively near-to-market systems, such as quantum-communication networks, ultra-sensitive cameras and quantum simulators that could help to design new materials. It will also look long term, pushing more-futuristic visions such as all-purpose quantum computers and high-precision sensors that fit into mobile phones.

Success will be judged by how well the flagship boosts industry take up of the technologies and seeds investment in the field, says Calarco: “If this doesn’t happen, it will be a failure. But everyone is very confident it will”.

Quantum-technology projects already exist in a few individual European Union countries, such as the UK Quantum Technologies Programme and the Netherlands’ QuTech initiative, notes Marco Genovese, a quantum physicist at the Italian National Institute of Metrological Research in Turin. But to reach commercial level in the near future, an EU-wide initiative is essential, he says. “At the moment, EU industry is still only marginally involved,” he says.

Europe’s graphene and brain-project flagships were announced with great fanfare in 2013, after a multiyear competition, but the latest initiative has had a much quieter birth. Calarco says that it was driven by an 18-month dialogue between the commission and a group of researchers who, at the organization’s request, produced the manifesto.

Not everyone is pleased with this approach. Choosing flagships on the basis of bilateral discussions and manifestos risks turning them into “a competition of lobbying, rather than of arguments evaluated objectively in a fair competition of scientific ideas”, says Adrian Ionescu, a nanoscientist at the Swiss Federal Institute of Technology in Lausanne. (Ionescu led an unsuccessful shortlisted project in the 2013 competition, called Guardian Angels for a Smarter Life, which would develop sensors to track environmental pollution and human health.) But the commission says that it is still running a separate consultation to identify candidates for future flagship projects, and that the quantum initiative does not prevent the launch of other flagships.

Genovese warns that the new project must be careful to avoid the problems faced by existing giant flagships, which included accusations of mismanagement and veering off course. “The building of the flagship must involve all the main research groups that have really significantly worked in the field through a bottom-up approach, and the concentration of power should be avoided,” he says.

The commission is set to announce more details at the Quantum Europe Conference in Amsterdam on 17–18 May, where the manifesto will be officially launched. ■

MICHAEL FANG/MARTINIS GROUP

HEALTH CARE

Drug firm seeks genome bounty

AstraZeneca aims to scan two million genomes in hunt for rare sequences linked to disease.

BY HEIDI LEDFORD

One of the world's largest pharmaceutical companies has launched a massive effort to compile genome sequences and health records from two million people over the next decade. In doing so, AstraZeneca and its collaborators hope to unearth rare genetic sequences that are associated with disease and with responses to treatment.

It's an unprecedented number of participants for this type of study, says Ruth March, vice-president and head of personalized health care and biomarkers at AstraZeneca, which is headquartered in London. "That's necessary because we're going to be looking for very rare differences among individuals."

To achieve that, AstraZeneca will partner with institutions including the Wellcome Trust Sanger Institute in Hinxton, UK, the Institute for Molecular Medicine Finland in Helsinki and Human Longevity, a biotechnology company

founded in San Diego, California, by genomics pioneer Craig Venter. AstraZeneca also expects to draw on data from 500,000 participants in its own clinical trials.

In doing so, AstraZeneca will be following a burgeoning trend in genetics research. For years, geneticists pursued common variations in human DNA sequences that are linked to complex diseases such as diabetes and heart disease. The approach yielded some important insights, but these common variations often accounted for only a small percentage of the genetic contribution to individual diseases.

Researchers are now increasingly focusing on the contribution of unusual genetic variants to disease. Combinations of these variants can hold the key to an individual's traits, says Venter.

AstraZeneca did not disclose exactly how much it would be investing — "hundreds of millions of dollars" over the course of ten years was all that Menelas Pangelos, executive vice-president of the company's innovative medicines

programme, would say. The company intends to use the data to inform drug development in all of its major disease areas, from diabetes to inflammation and cancer, says March.

Genomicists have long promised that their field would revolutionize drug development, says David Goldstein, a geneticist at Columbia University in New York City who advises AstraZeneca. Now, he says, "we finally have really turned a corner and genomics really will now become central in drug development". ■

CORRECTION

The News Feature 'Monkey Kingdom' (*Nature* **532**, 300–302; 2016) wrongly affiliated Erwan Bezard with INSERM — he is actually director of the Institute of Neurodegenerative Diseases at the University of Bordeaux. It also referred to Liping Zhang instead of Liping Wang.

LISTENING FOR LANDSLIDES

A year after a devastating earthquake triggered killer avalanches and rock falls in Nepal, scientists are wiring up mountainsides to forecast hazards.

BY JANE QIU

Kodari is a ghost town on an empty Nepalese highway that cuts through some of the steepest slopes of the Himalayas. One year after the magnitude-7.8 Gorkha earthquake killed nearly 9,000 people, the once-buzzing trade centre looks like a battlefield where armies of giants once waged war. The road is littered with rusting cars and trucks smashed into bizarre shapes. Massive boulders rest on the wreckage of homes.


“It’s a good example of building a town in the wrong place,” says Kristen Cook, a geologist at the German Research Centre for Geosciences (GFZ) in Potsdam, as she climbs over the rubble from one of the landslides that crushed the town. The Arniko Highway, which runs through Kodari, is no stranger to such calamities, especially in the monsoon season. “It was in frequent repair and closure even before the earthquake,” says Shanmukesh Amatya, landslide-division chief at Nepal’s Department of Water Induced Disaster Prevention in Kathmandu. “The problem now is overwhelming.”

The highway is not the only thing that keeps Amatya awake at night. The earthquake unleashed more than 10,000 landslides that blocked rivers and damaged houses, roads and other key pieces of infrastructure across the country. And the destruction didn’t stop with the shaking. The hilly terrain, severely weakened by the quake, is now more likely

to slip after strong rains and aftershocks — a legacy that is likely to endure for years. During the most recent monsoon, the area affected by landslides was about ten times greater than usual.

“It’s a real problem for reconstruction,” says Tara Nidhi Bhattarai, a geologist at Tribhuvan University in Kathmandu and chief

SAMIR AMIR JUNG THAPA/EPA



Locals peer at
a landslide in
Langtang, Nepal.

scientist of Nepal's National Reconstruction Authority — an agency established last year to manage the recovery efforts. “What are the safe places to rebuild, in a landscape that is evolving?”

To answer that, geoscientists are wiring up the mountains in Nepal and other seismically active countries. By monitoring how hillsides

evolve, researchers are learning why strong shaking weakens a slope and makes it more prone to give way during aftershocks or rainstorms. The lessons from such studies could help to pinpoint when and where the side of a mountain will collapse.

NATURE.COM

To hear more about
landslides in Nepal:
go.nature.com/fmiidu

The significance goes beyond quake recovery. Himalayan nations are facing increasing risks from landslides because of deforestation, road construction, population growth and other changes that have pushed people to live in hazardous locations. Climate change may exacerbate the problem by melting glaciers and triggering increasingly extreme rainfall.

“There is a pressing need to monitor the risks in the long run,” says Amatya. “A nationwide early-warning system is long overdue.”

BIRD'S-EYE VIEW

A crowd eagerly looks on as Cook flies a drone through the skies near Listi, a small village perched on a mountainside above the Arniko Highway. With its four propellers, the little robot zips over landslide scars that run down from the ridge like gigantic frozen waterfalls.

A camera and other sensors on the drone provide data that let Cook build a 3D reconstruction of the landscape. She started the work last October and will take measurements every few months over the next few years. By scanning as many landslide-inflicted areas as possible, she says, “we will be able to trace how they change over time and what’s the effect of monsoons”.

Such measurements of the surface will complement studies that track what’s happening underground. Not far from Cook is her colleague Christoff Andermann, another GFZ geologist, who is performing maintenance on a broadband seismometer, a device that measures shaking across a wide range of frequencies. Last June, the GFZ team installed a dozen such instruments, along with weather stations and river-flow sensors, across 50 square kilometres of landslide-riddled terrain.

Seismometers are a relatively new addition to landslide studies by the GFZ researchers and their colleagues. They started using the sensors only after an accidental discovery. In 2003, a set of seismic stations installed in Nepal to study deep structures in Earth’s crust picked up high-frequency noise from nearby rivers and shifting slopes. Arnaud Burtin, a seismologist now at the Earth Physics Institute in Paris, noticed a series of peaks in that noise before a debris flow in central Nepal that killed 45 people. He and his colleagues went on to identify¹ 46 debris flows from seismograms taken during that monsoon season. By comparing the data with information from weather stations, the team also determined how much rainfall was required to trigger slides.

Researchers have typically used satellite imagery or aerial photography to track landscape changes on a large scale, but these methods have relatively poor temporal resolution because images are taken days or months apart. Seismometers take snapshots hundreds of times per second, so they are ideal for monitoring slopes for instability, says Colin Stark, a geologist at the Lamont-Doherty Earth Observatory in Palisades, New York, who studies monster landslides using global seismic networks. When seismometers are placed strategically, he says, it’s also possible to precisely locate the source of seismic signals in a large area.

“Until recently, we had little idea why landslides are more likely to happen after an earthquake or how the slopes recover over time,” says Stark. But work over the past decade has revealed that cracks produced by an earthquake can boost the shaking in future shocks. Unpublished results from seismic stations, for example, show that on fractured slopes, ground motion can be up to 30 times what is measured in neighbouring, undamaged areas, says Jeffrey Moore, a geophysicist at the University of Utah in Salt Lake City. This means that minor after-shocks could trigger unexpected levels of landslides in damaged slopes that did not fail in the main shock, he says.

In some cases, the increased sensitivity can last for decades. A study² of a magnitude-7.4 earthquake in New Zealand in 1968 found that the quake triggered more landslides than expected in places that had been affected by a magnitude-7.8 shock 21 kilometres away and nearly 4 decades before.

Quake-stricken hills also have an increased sensitivity to rainfall, says Niels Hovius, a GFZ geologist who is leading the Nepal study. He and his colleagues have found³ that after the magnitude-7.6 ChiChi earthquake that hit Taiwan in 1999, the rate of rainfall-triggered landslides in the affected area jumped by a factor of 22. “The government cleared up the mess and rebuilt, but the same happened again a couple of years later,” he says. If scientists can develop greater insight into the mechanisms that control slope behaviour after an earthquake, that could help authorities to make better decisions about rebuilding.

By analysing records after the ChiChi quake and three others with similar depths and slip mechanisms, Hovius and his colleagues also found³ that it took up to four years for landslide rates to return to pre-quake levels at those sites.

In follow-up work, the team mined data from seismometers installed before ChiChi hit. The instruments were near roads, which

“Until recently, we had little idea why landslides are more likely to happen after an earthquake.”

made it possible to study subsurface properties by measuring how traffic vibrations travel through the ground. They found that the speed of seismic waves dropped markedly immediately after the quake. The velocities then recovered gradually, following roughly the same trajectory as the decline in landslide rates, says Odin Marc, a geologist at the GFZ, who presented the results last week at a meeting of the European Geosciences Union in Vienna. Over the same period, there were frequent, small surface displacements — presumably caused

by the slow, creeping movement of Earth’s crust after an earthquake, a process known as post-seismic deformation.

The researchers suspect that subsurface materials are packed together tightly before the earthquake, like beads in a box. Strong ground-shaking causes the granular mass to expand, opening up holes and cracks that make the ground less dense. “This is why seismic waves travel at reduced speeds,” says Hovius. Post-seismic deformation causes the openings to fill in and the subsurface sediments to become compact once more. “It’s an internal healing process of the landscape,” he says.

Data collected after the Gorkha earthquake support that. Preliminary results show that seismic-wave velocities close to the surface declined sharply after the shock — and the volume of water flowing through rivers increased by 50%. That backs up the idea that the quake opened holes and fractures in the subsurface, which then allowed groundwater to leak more freely through the cracks, says Andermann, who has been monitoring river flows and sediment transport in the region for the past decade.





BORJA SANCHEZ TRILLO/GETTY

**Landslides
devastated villages
near Kodari, Nepal.**

At high-risk sites in Nepal, researchers are combining seismological and other techniques to watch for signs that mountainsides are growing restless. On the steep slope facing Listi, the earthquake caused the lower part of the ridge to subside, resulting in a 5-metre opening that skirts the mountain for about 2 kilometres. This gigantic crack and many smaller ones nearby pose a serious threat to downslope settlements, says Amod Dixit, executive director of Nepal's National Society for Earthquake Technology (NEST) in Kathmandu. "They must be closely monitored."

Last August, Nick Rosser, a geologist at Durham University, UK, and his colleagues installed a series of instruments at ten locations across the slope — including strain meters to monitor changes in the cracks, accelerometers to measure ground vibration, and rain gauges. The data are relayed to a server at NEST, letting researchers track in real time whether the cracks are opening or contracting and how they respond to rainfall.

Although it is not yet a fully fledged early-warning system, the set-up can identify signs of major deformation that could cause the slope to fail. Thankfully, says Rosser, "the cracks are not growing at the moment". Settlements will be alerted to any impending danger, he adds.

The researchers are using information from the field and from lab experiments on slope materials to try to determine what kind of ground deformation and rainfall would cause landslides. "This is crucial for setting the criteria for triggering an alert," he says.

The Durham sensors are within the area covered by the GFZ seismic array, so the teams will pool their field data. Together with satellite imagery and other measurements, this information will provide unprecedented insight into how the mountains are changing and what kind of danger this might pose to communities there, they say.

At Listi, Cook is worried about a massive pile of debris that the drone has located high above the valley. The earthquake loosened a huge amount of rock and soil, but most did not make it all the way to the bottom. "They are just sitting there on the hillside," says Cook, pointing to a mass on her remote-control screen. The materials could all come down in heavy rain — as some did during the last monsoon. "They are time bombs waiting to explode." ■

Jane Qiu is a writer in Beijing. Her trip to Nepal was supported by a grant from the Pulitzer Center on Crisis Reporting.

1. Burtin, A., Bollinger, L., Cattin, R., Vergne, J. & Nábělek, J. L. J. *Geophys. Res.* **114**, F04009 (2009).
2. Parker, R. N. et al. *Earth Surf. Dynam.* **3**, 501–525 (2015).
3. Marc, O., Hovius, N., Meunier, P., Uchida, T. & Hayashi, S. *Geology* **43**, 883–886 (2015).
4. Kargel, J. S. et al. *Science* **351**, aac8353 (2016).
5. Lacroix, P. *Earth Planets Space* **68**, 46 (2016).

Such findings suggest a way to predict landslides. Looking back over their data, the researchers were able to identify peaks of seismic signals in the run-up to a major landslide last July. "These precursors represent a sequence of processes that culminated in the failure," says Hovius. "There was a systematic increase in the rate at which these precursor activities occurred, until the whole topography collapsed."

The GFZ team also found that seismic waves travel through the subsurface more quickly when the slope is drenched and pore spaces are filled with water. "We can see how quickly the effects of rainfall propagate into and through the subsurface" using seismic sensors, says Hovius. This effectively maps groundwater flow, a key factor in the strength of hillsides. With the seismic data, researchers can model the physics of slope stability and monitor changes in ground properties that might precipitate a landslide.

NEAR-ATOMIC BLAST

In the village of Langtang in northern Nepal, a pile of rubble 60 metres deep provides ample incentive to improve landslide forecasts. During the earthquake last year, a mixture of ice and rock crashed down several kilometres onto the valley floor — landing with an impact that released half as much energy as the Hiroshima atomic bomb⁴. The slide buried

Langtang and nearby villages, leaving nearly 400 people dead or missing.

Research groups have been racing to understand where the avalanche began and whether the area is still at risk. One study⁵ found 5 initiation sites between altitudes of 6,800 and 7,200 metres, along a 3-kilometre ridge where the earthquake shook up snow and glaciers. These swept down the slope, picking up rocks as they went.

Roughly 7 million cubic metres of debris filled the bottom of the valley, and another 10 million cubic metres still rest precariously on slopes more than 5,000 metres above sea level. A year after the quake, the sounds of falling rocks and shifting slopes frequently echo through the valley — a reminder of the remaining hazard.

The Langtang case shares features with increasingly common rock avalanches in high mountains in Alaska and the Alps, says Marten Geertsema, a glaciologist with the British Columbia Ministry of Forests and Range in Prince George, Canada. In all these places, glaciers are quickly retreating, leaving rocky hillsides exposed and prone to failure. And warming at high elevations may cause frozen bedrock to thaw, he says, making it more permeable to melt water and weakening the rocks. "Climate change might have primed the landscape for the devastation."



A CRISPR VISION

Emmanuelle Charpentier spent years moving labs and relishing solitude. Then the co-discovery of CRISPR–Cas9 explosively changed her life.

BY ALISON ABBOTT

Emmanuelle Charpentier's office is bare, save for her computer. Her pictures, still encased in bubble wrap, are stacked in one corner, and unpacked cardboard boxes stuffed with books and papers are lined up in the adjacent room. But across the corridor, her laboratory is buzzing with activity. When Charpentier moved to Berlin six months ago, she had her science up and running within weeks, but decided that the rest could wait. "We were all determined to get the research going as fast as possible," she says, leaning forward from her still-pristine office chair.

Charpentier's workspace is a fitting reflection of her scientific life — one in which she always seems to be moving while keeping science on the go. Now 48, she has climbed her way up the academic ladder by way of 9 different institutes in 5 different countries over the past 20 years. "I always had to build up new labs from scratch, on my own," she says. Her eureka

moments have occurred amid packing boxes and, after years on short-term grants, she was 45 before she was able to employ her own technician. "She's so resourceful, she could start a lab on a desert island," says Patrice Courvalin, her PhD supervisor at the Pasteur Institute in Paris.

The itinerant lifestyle doesn't seem to have hampered the microbiologist as she has carefully dissected the systems by which bacteria control their genomes. Charpentier is now acknowledged as one of the key inventors of the gene-editing technology known as CRISPR–Cas9, which is revolutionizing biomedical researchers' ability to manipulate and understand genes. This year, she has already won ten prestigious science prizes, and has officially taken up a cherished appointment as a director of the Max Planck Institute for Infection Biology in Berlin. The gene-therapy company that she co-founded in 2013, CRISPR Therapeutics, has become one of the world's most richly financed preclinical biotech companies, and she is in the middle of a high-profile patent dispute over the technology. Last September, Charpentier's phone kept on ringing. Journalists from around the world were trying to reach her, thinking — prematurely, as it turned out — that the imminent announcement of the 2015 Nobel prizes might well include her.

The academic limelight is not a comfortable place for Charpentier, which is why she remains the least well known member of the small international group tipped for the 'CRISPR Nobel', if it arrives. "Jean-Paul Sartre, the French philosopher, warned that winning prizes turned

PETER STEFFEN/DPA/PA

Emmanuelle Charpentier: a key inventor of the gene-editing technology CRISPR–Cas9.

paper¹ in *Nature* that reveals the mechanism of a CRISPR system that might prove even more efficient than CRISPR–Cas9.

Colleagues who know Charpentier well describe her as intense, modest and driven. “She’s a tiny person, with a very strong will — and she can be pretty stubborn,” says Rodger Novak, who was a postdoctoral researcher with her in the 1990s and is now chief executive of CRISPR Therapeutics. As Courvalin sees it, “She is like a dog with a bone — tenacious.”

MEDICAL MISSION

Small and slight, with eyes so dark that they seem black, Charpentier looks as restless as she evidently is. Growing up in a small town near Paris, she had a clear idea from the start of what she wanted in life: to do something to advance medicine. A visit to an aunt, a missionary who was living in an old convent, set her dreaming of being able to do this “in a lovely setting, where you can be a bit alone with yourself”.

Her socially engaged parents, she says, supported her ideas without guiding her in any direction. She pursued piano and ballet — but her leaning towards medicine eventually flowered into studies in life sciences. As an undergraduate at Pierre and Marie Curie University in Paris, she decided to do her PhD at the nearby Pasteur Institute, which was gaining a strong reputation in basic research and had a programme on antibiotic resistance that she wanted to join. Her PhD project involved analysing pieces of bacterial DNA that move around the genome and between cells, allowing drug resistance to be transferred.

Her years at the Pasteur Institute were formative. Her department in the historic institution was “young and fun,” she says. She loved to study at the old St Geneviève library close to Notre-Dame Cathedral, happily isolated in the triangle of light from the green-topped desk lamps. “I realized I had found my environment,” she says. Her ambition was to lead a lab at the Pasteur, and she decided that this would require a postdoc period abroad to gain expertise. “I was a typical French student of the 1990s — I imagined that after a short excursion I would work the rest of my life at home.”

Charpentier sent out 50 or so exploratory letters to labs in the United States, and got a postbag full of offers in reply. She chose a position with microbiologist Elaine Tuomanen at the Rockefeller University in New York City to work on the pathogen *Streptococcus pneumoniae*. This microbe, which is a major cause of pneumonia, meningitis and septicaemia, has a particularly free-wheeling relationship with mobile genetic elements, shifting them about its genome while maintaining its vicious pathogenicity. Tuomanen’s lab had priority access to its recently sequenced genome, offering the tantalizing prospect of discovering where these elements were landing and what happened when they did.

Charpentier carried out a stream of painstaking experiments to work out how the pathogen monitors and controls such elements, and contributed to a study identifying how the pathogen acquires resistance to vancomycin, an antibiotic of last resort². She had set out for New York with some trepidation but, absorbed in her work, was surprised to find that she wasn’t homesick. When Tuomanen moved her lab to Memphis, Tennessee, Charpentier wanted to stay, so she found a home in the lab of skin-cell biologist Pamela Cowin at New York University School of Medicine, where she also had the opportunity to learn about mammalian genes through working on mice.

Cowin remembers Charpentier as her first postdoc who did not need looking after. “She just ran with the programme,” she says. “She was driven, meticulous, precise and detail-oriented” — as well as a rather quiet, private person. Charpentier soon discovered that genetically modifying mice was a lot harder than manipulating bacteria. She spent two years on the project and emerged with a paper on the regulation of hair growth, a solid grounding in mammalian genetics and a strong

you into an institution — I am just trying to keep working and keep my feet on the ground,” she says. She seems to be succeeding, this week publishing a

desire to develop better tools for genetic engineering.

After another postdoc in New York, Charpentier knew that her next step needed to be complete independence — and a move back to Europe. Her time in the United States had taught her that she was European rather than solely French, and she chose Vienna. She arrived at the university there in 2002, and spent the next seven years running a small lab that was precariously dependent on short-term grants. “I had to survive on my own,” she says. Nevertheless, “I had in mind to understand how every biochemical pathway in a bacterium was regulated.” It was an exciting time scientifically, with the importance of small RNA molecules in regulating genes being revealed, and she embarked on many different projects on various bacteria — possibly too many, she admits, but she kept winning the grants. She discovered an RNA that controls the synthesis of a class of molecules that are important for virulence in the bacterium *Streptococcus pyogenes*³.

It was in Vienna that Charpentier first found herself thinking about CRISPR. In the early 2000s, this was a niche area: only a handful of microbiologists were paying attention to the newly discovered, curiously patterned stretch of DNA called CRISPR in the genome of some bacteria, where it serves as part of a defence system against viruses. By copying part of an invading virus’ DNA and inserting it into that stretch, bacteria are able to recognize the virus if it invades again, and attack it by cutting its DNA. Different CRISPR systems have different ways of organizing that attack; all of the systems known at the time involved an RNA molecule called CRISPR RNA.

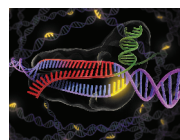
Charpentier was interested in identifying sites in the genome of *S. pyogenes* that made regulatory RNAs — and found that bioinformatics took her only so far. So she forged a collaboration with molecular microbiologist Jörg Vogel, then a junior group leader at the Max Planck Institute for Infection Biology, who was developing methods for large-scale mapping of RNAs in a genome. He agreed to map *S. pyogenes* — and by 2008 he had sequences of all of the small RNAs generated by the bacterium.

The first thing that the researchers noticed was a super-abundance of a novel small RNA that they called trans-activating CRISPR RNA (tracrRNA). From its sequence and position on the genome — it was at a location that Charpentier’s bioinformatics had predicted as being close to the CRISPR site — they realized that it was highly likely to be involved in a CRISPR system that had not previously been described. Charpentier and her colleagues began a long series of experiments to explore this system, identifying that it had just three components — tracrRNA, CRISPR RNA and the Cas9 protein. This was a surprise: “Other CRISPR systems involved just one RNA and many proteins, and no one had really thought that two RNAs might be involved,” says Charpentier. The system was so exceptionally simple that she realized that it might one day be harnessed as a powerful genetic engineering tool. If the components could be controlled, it might provide the long-sought ability to find, cut and potentially alter DNA at a chosen, precise site in a genome.

But how exactly was this CRISPR system working? Charpentier suspected that the two RNAs might actually interact with each other to guide Cas9 to a particular DNA sequence in the virus. The concept

was radical; that type of teamwork is routine for proteins, but not for RNAs. But Charpentier “always looked for the unexpected rather than the expected in a genome,” says Tuomanen. “She is a very counter-culture person.” Charpentier remembers that it was hard to persuade any of her young students to follow up her intuition and perform the key experiment to test whether the two RNAs might interact, but eventually a masters student at the University of Vienna, Elitza Deltcheva, volunteered.

“SHE’S SO RESOURCEFUL,
SHE COULD START A LAB
ON A DESERT ISLAND.”



NATURE.COM

For more of Nature’s coverage of CRISPR, see: nature.com/crispr

By then, it was June 2009, and Charpentier was again on the move. She had never felt completely at home in Vienna, where she says the grandiose architecture oppressed her. And she knew that she had to find more security and support. “At this time in my career, I needed the luxury of being able to focus on finalizing a big, cool story,” she says. She took a position at the newly created, well provisioned Umeå Centre for Microbial Research in northern Sweden. The pretty, human-scale architecture of the old town made her feel comfortable, and she even learned to like the long, dark winters, which made her lose the feeling of time, allowing an even greater focus on work.

In summer 2009, she was still commuting between Austria and Sweden when Deltcheva called her in Umeå at 8 p.m. to tell her that the experiments had worked. “I was very, very happy,” Charpentier says. But she told no one. Vogel says that it was “a very intense time”. He recalls getting a call from Charpentier one night that August when he was driving on a country road outside Berlin. “I stood on the kerbside for ages while we discussed when would be the right time to publish, because by then we had actually got the story.”

They both knew that this discovery was going to be a game-changer, but both were afraid of being scooped if word of the system they had stumbled on got out. To make sure that publication would not be drawn out by referees’ queries, they worked doggedly and silently for more than a year to cover as many bases as they could think of before submitting to *Nature*⁴.

Charpentier was unknown in the then-small CRISPR world. She presented the work for the first time in October 2010 at a CRISPR meeting in Wageningen, the Netherlands, a few weeks after submitting it for publication. “It was a highlight of the meeting — a beautiful story that was extremely unexpected and came right out of the blue,” says microbiologist John van der Oost of Wageningen University, who organized the meeting. Charpentier didn’t mind being an outsider. “I have never really wanted to be part of a cosy scientific community,” she says. And she was already thinking ahead to the next step — how this neat dual-guide RNA system actually led to cleavage of DNA.

At a 2011 American Society for Microbiology conference in San Juan, Puerto Rico, she met structural biologist Jennifer Doudna of the University of California, Berkeley. Doudna was immediately charmed. “I loved her intensity, which was apparent from the moment I met her,” she says. They began a collaboration that swiftly led to the second key discovery showing how Cas9 cleaved DNA⁵. With the mechanism elucidated, researchers went on to show that the system could indeed be adapted to make targeted cuts in a genome and to modify a sequence. The technique has since been embraced by labs around the world.

Charpentier, meanwhile, made two decisions. The first was in deference to her original ambition to do something to advance medicine. She contacted Novak, who was by then working at the pharmaceutical firm Sanofi in Paris, with the intention of co-founding a company to exploit the methodology for human gene therapy. CRISPR Therapeutics, based in Cambridge, Massachusetts, and Basel, Switzerland, was born in November 2013 with a third co-founder, Shaun Foy, and Charpentier remains chair of its scientific advisory board.

The second decision was in deference to her ambition to fully dedicate

“THE SCIENTIST THAT I AM GOT ME HERE, AND THAT IS THE SCIENTIST THAT I WANT TO REMAIN.”



Jennifer Doudna (left) and Emmanuelle Charpentier receive the Breakthrough Prize in Life Sciences in November 2014.

her time to basic research in gene regulation. For this she wanted a permanent post, with more institutional support.

In 2013, she moved to Germany to become a professor at the Hanover Medical School and a department chief at the Helmholtz Centre for Infection Research in nearby Braunschweig, where she finally got her own technicians and built up a lab of 16 PhD students and postdocs. Just over two years later, she was recruited by the Max Planck Institute in Berlin. Now she has generous technical and institutional support, and her labs are in the elegant, nineteenth-century campus of the Charité teaching hospital, an environment she can relax in. Maybe in a few years, she says, she’ll even find a few moments for reading philosophy.

But right now, fame and prize-winning leave little time for that. She values the recognition, engaging fully with the publicity activities that each prize requires — but notes anxiously that on average, each takes two full days from work. She declines to discuss the high-profile

and rather complicated patent dispute between herself — alongside Doudna and Berkeley — and the Broad Institute of MIT and Harvard in Cambridge, Massachusetts. She leaves that to the patent lawyers, who are currently arguing it out.

Her focus is still on research, and her latest paper¹ — an elaboration of a CRISPR system that is even simpler than CRISPR-Cas9 — was once again finalized in the middle of a lab move. The work shows that a protein called Cpf1 can do the jobs of both tracrRNA and the Cas9 protein — “a very important contribution,” says van der Oost, and part of a flurry of recent studies on this system^{6,7}. But Charpentier is keen not to be defined by CRISPR, which is just one of five themes in her lab; others include the mechanisms by which pathogens interact with host immune cells and the molecular complexes that regulate the behaviour of bacterial chromosomes.

Reflecting back, she feels that her life has been tougher than it need have been. She notes that now there are more sources of major grants to help young investigators to start their own independent labs. And although her goals to further medicine and improve genetic-engineering tools have been met, her ambitions have not waned. “I haven’t changed, and I won’t change,” she says. “The scientist that I am got me here, and that is the scientist that I want to remain.”

But some things have changed. Charpentier is not an outsider any more: she is an established member of the rapidly expanding CRISPR community and is inundated with invitations to give talks. Her mischievous ambition, however, is to show up at a CRISPR meeting and report the discovery of something entirely different, but equally important. She has a few things up her sleeve, she says. ■

Alison Abbott is *Nature’s* senior European correspondent.

1. Fonfara, I. *et al. Nature* **532**, 517–521 (2016).
2. Novak, R., Henriques, B., Charpentier, E., Normark, S. & Tuomanen, E. *Nature* **399**, 590–593 (1999).
3. Mangold, M. *et al. Mol. Microbiol.* **53**, 1515–1527 (2004).
4. Deltcheva, E. *et al. Nature* **471**, 602–607 (2011).
5. Jinek, M. *et al. Science* **337**, 816–821 (2012).
6. Zetsche, B. *et al. Cell* **163**, 759–771 (2015).
7. Dong, D. *et al. Nature* **532**, 522–526 (2016).

COMMENT

TECHNOLOGY Why did the Soviets lose the Internet race? **p.438**



HISTORY Theodore Roosevelt's love of nature launched national parks **p.440**

CAREERS Don't rob postdocs of rights to boost lab productivity **p.441**

OBITUARY R. McNeill Alexander, animal-biomechanics pioneer, remembered **p.442**

ANEK SUWANNAPHOOM



Refineries use huge amounts of thermal energy to process crude oil.

Seven chemical separations to change the world

Purifying mixtures without using heat would lower global energy use, emissions and pollution — and open up new routes to resources, say **David S. Sholl** and **Ryan P. Lively**.

Most industrial chemists spend their days separating the components of large quantities of chemical mixtures into pure or purer forms. The processes involved, such as distillation, account for 10–15% of the world's energy consumption^{1,2}.

Methods to purify chemicals that are more energy efficient could, if applied to the US petroleum, chemical and paper manufacturing sectors alone, save 100 million tonnes of carbon dioxide emissions and US\$4 billion in energy costs annually³ (see 'Cutting costs').

Other methods would enable new sources of materials to be exploited, by extracting metals from seawater, for example.

Unfortunately, alternatives to distillation, such as separating molecules according to their chemical properties or size, are underdeveloped or expensive to scale up. Engineers in industry and academia need to develop better and cheaper membranes and other ways to separate mixtures of chemicals that do not rely on heat.

Here, we highlight seven chemical

separation processes that, if improved, would reap great global benefits. Our list is not exhaustive; almost all commercial chemicals arise from a separation process that could be improved.

SEVEN SEPARATIONS

Hydrocarbons from crude oil. The main ingredients for manufacturing fossil fuels, plastics and polymers are hydrocarbons. Each day, refineries around the world process around 90 million barrels of crude ►

► oil — roughly 2 litres for every person on the planet. Most do so using atmospheric distillation, which consumes about 230 gigawatts (GW) per year globally³, equivalent to the total energy consumption of the United Kingdom in 2014 or about half that of Texas. In a typical refinery, 200,000 barrels per day of crude oil are heated in 50-metre-tall columns to liberate thousands of compounds according to their boiling points. Light gases emerge at the cool top (at around 20°C); progressively heavier fluids leave at lower and hotter points (up to 400°C).

Finding an alternative to distillation is difficult because crude oil contains many complex molecules, some with high viscosities, and myriad contaminants, including sulfur compounds and metals such as mercury and nickel. It is feasible in principle to separate hydrocarbons according to their molecular properties, such as chemical affinity or molecular size. Membrane-based separation methods, or other non-thermal ones, can be an order of magnitude more energy efficient than heat-driven separations that use distillation. But little research has been done.

Researchers need to find materials that are capable of separating many families of molecules at the same time, and that work at the high temperatures needed to keep heavy oils flowing without becoming blocked by contaminants.

Uranium from seawater. Nuclear power will be crucial for future low-carbon energy generation. Although the trajectory of the nuclear industry is uncertain, at current consumption rates, known geological reserves of uranium (4.5 million tonnes) may last a century⁴. More than 4 billion tonnes of uranium exist in seawater at part-per-billion levels.

Scientists have sought ways to separate uranium from seawater⁴ for decades. There are materials capable of capturing uranium, such as porous polymers containing amidoxime groups. But these molecular ‘cages’ also capture other metals, including vanadium, cobalt and nickel.

Chemists need to develop processes to remove these metals while purifying and concentrating uranium from seawater. In 1999–2001, Japanese teams captured around 350 grams of uranium using an adsorbent fabric⁴. Starting up a new nuclear power plant requires hundreds of tonnes of uranium fuel, so the scale of these processes would need to be vastly increased. In particular, efforts to reduce costs for adsorbent materials are needed.

Similar technologies could capture other valuable metals⁴, such as lithium, which is used in batteries. The quantity of lithium

dissolved in the oceans is ten times larger than that in known land-based resources; the limited size of the latter may become a long-term barrier to energy storage.

Alkenes from alkanes. Manufacturing plastics such as polyethene and polypropene requires alkenes — hydrocarbons such as ethene and propene, also known as olefins. Global annual production of ethene and propene exceeds 200 million tonnes, about 30 kilograms for each person on the planet. The industrial separation of ethene from ethane typically relies on high-pressure cryogenic distillation at temperatures as low as –160°C. Purification of propene and ethene alone accounts for 0.3% of global energy use, roughly equivalent to Singapore’s annual energy consumption.

As with crude oil, finding separation systems that do not require changes from one phase to another could reduce by a factor of ten the energy intensity of the process (energy used per unit volume or weight of product), and offset carbon emissions by a similar amount⁵. For example, porous carbon membranes are being developed that can separate gaseous alkenes and alkanes (also called paraffins) at room temperature and at mild pressures (less than 10 bar)⁶. But these cannot yet produce the more than 99.9% pure alkenes needed for chemicals manufacturing.

In the short term, ‘hybrid’ separation techniques might help — membranes can be used for bulk separation and cryogenic distillation for ‘polishing’ the product. Such approaches would reduce the energy intensity of alkene production by a factor of 2 or 3, until membranes become good enough to replace distillation entirely. A major hurdle is scaling up the membranes — industry might require surface areas of up to 1 million square metres. Deployment on this scale will require new manufacturing methods as well as advances in materials’ properties.

Greenhouse gases from dilute emissions. Anthropogenic emissions of CO₂ and other hydrocarbons, such as methane released from refineries and wells, are key contributors to global climate change. It is expensive and technically difficult to capture these gases from dilute sources such as power plants, refinery exhausts and air.

Liquids such as monoethanolamine react readily with CO₂, but because heat must be applied to remove CO₂ from the resulting liquid, the process is not economically viable for power plants. If the approach was applied to every power station in the United States, CO₂ capture could cost 30% of the country’s growth in gross domestic product each year⁷. Cheaper methods for capturing CO₂ and hydrocarbon emissions with minimal energy costs need to be developed.

A complicating factor is deciding what to do with the purified product. CO₂ could be used in a crude-oil production method known as enhanced oil recovery, or in vertical farming and as chemical and biorefinery feedstocks. But human activities emit so much of the gas⁸ that in practice much of it will need to be stored long term in underground reservoirs, raising other issues.

Rare-earth metals from ores. The 15 lanthanide metals, or rare-earth elements, are used in magnets, in renewable-energy technologies and as catalysts in petroleum refining. Compact fluorescent lamps use europium and terbium, for example, and catalytic converters rely on cerium. Producing rare earths economically is a problem of separation, not availability. Despite their name, most of the elements are much more plentiful in Earth’s crust than gold, silver, platinum and mercury. Unfortunately, rare earths are found in trace quantities in ores and are often mixed together because they are chemically similar.

Separation of rare earths from ores



High-capacity (HiCap) polymers can separate metals such as uranium from solution.

ORNL

requires mechanical approaches (such as magnetic and electrostatic separation) and chemical processing (such as froth flotation). These are inefficient: they must contend with the complex compositions of mined ores, use large volumes of chemicals, and produce lots of waste and radioactive by-products. Improvements are sorely needed.

The recycling of rare earths from discarded products is increasing. Bespoke processes could be designed because the chemical and physical compositions of the products are well defined. A variety of metallurgical and gas-phase extraction methods have been explored, but recycled rare earths are not yet part of most supply chains^{9,10}. Research is needed to reduce the ecological impact of key items containing rare earths over their whole life cycle.

Benzene derivatives from each other. The supply chains of many polymers, plastics, fibres, solvents and fuel additives depend on benzene, a cyclic hydrocarbon, as well as on its derivatives such as toluene, ethylbenzene and the xylene isomers. These molecules are separated in distillation columns, with combined global energy costs of about 50 GW, enough to power roughly 40 million homes.

The isomers of xylene are molecules with slight structural differences from each other that lead to different chemical properties. One isomer, para-xylene (or *p*-xylene), is most desirable for producing polymers such as polyethylene terephthalate (PET) and polyester; more than 8 kilograms of *p*-xylene is produced per capita each year in the United States. The similar size and boiling points of the various xylene isomers make them difficult to separate by conventional methods such as distillation.

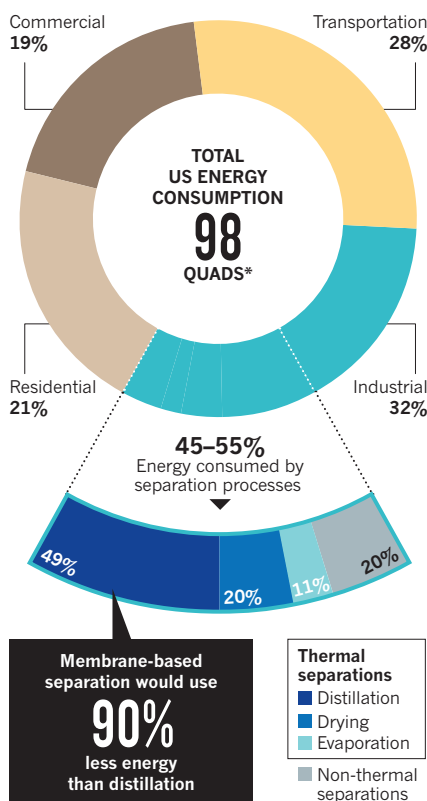
Advances in membranes or sorbents could reduce the energy intensity of these processes. As for other industrial-scale chemical processes, implementing alternative technologies for separating benzene derivatives will require that their viability be proved on successively larger scales before commercial implementation. Constructing a chemical plant can cost US\$1 billion or more, so investors want to be sure that a technology will function before building new infrastructure.

Trace contaminants from water. Desalination — whether through distillation or membrane filtration — is energy and capital intensive, making it unfeasible in many dry areas. Distillation is not the answer: thermodynamics defines the minimum amount of energy needed to generate potable water from seawater, and distillation uses 50 times more energy than this fundamental limit.

Reverse-osmosis filtration, a process that applies pressure across a membrane to salty water to produce pure water, requires only

CUTTING COSTS

Chemical separations account for about half of US industrial energy use and 10–15% of the nation's total energy consumption. Developing alternatives that don't use heat could make 80% of these separations 10 times more energy efficient.



*A quad is a unit of energy equal to 10^{15} British Thermal Units (1 BTU is about 0.0003 kilowatt-hours).

25% more energy than the thermodynamic limit⁵. But reverse-osmosis membranes process water at limited rates, requiring large, costly plants to produce a sufficient flow. Reverse osmosis of seawater is already done on commercial scales in the Middle East and Australia. But the practical difficulties of handling more-polluted water — including corrosion, biofilm formation, scaling and particulate deposition — mean that expensive pretreatment systems are also needed.

Developing membranes that are more productive and resistant to fouling would drive down the operating and capital costs of desalination systems to the point that the technique is commercially viable for even highly polluted water sources.

NEXT STEPS

Academic researchers and policymakers should focus on the following issues.

First, researchers and engineers must consider realistic chemical mixtures. Most academic studies focus on single chemicals and infer the behaviour of mixtures using this information. This approach risks missing phenomena that occur only in

chemical blends, and ignores the role of trace contaminants. Academics and leaders in industrial research and development should establish proxy mixtures for common separations that include the main chemical components and common contaminants.

Second, the economics and sustainability of any separation technology need to be evaluated in the context of a whole chemical process. Performance metrics such as cost per kilogram of product and energy use per kilogram should be used. The lifetime and replacement costs of components such as membrane modules or sorbent materials need to be factored in.

Third, serious consideration must be given early in technology development to the scale at which deployment is required. Physical infrastructure such as academic and industrially operated test beds will be needed to take new technologies from the lab to pilot scales so that any perceived risk can be reduced. Managing this will require academia, government agencies and industry partners to collaborate.

Fourth, current training of chemical engineers and chemists in separations often places heavy emphasis on distillation. Exposure to other operations — such as adsorption, crystallization and membranes — is crucial to develop a work force that is able to implement the full spectrum of separations technologies that the future will require. ■

David S. Sholl and Ryan P. Lively are professors in the School of Chemical & Biomolecular Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA.
e-mail: david.sholl@chbe.gatech.edu

1. Oak Ridge National Laboratory. *Materials for Separation Technologies: Energy and Emission Reduction Opportunities* (2005).
2. Humphrey, J. & Keller, G. E. *Separation Process Technology* (McGraw-Hill, 1997).
3. US Dept. Energy Advanced Manufacturing Office. *Bandwidth Study on Energy Use and Potential Energy Saving Opportunities in U.S. Petroleum Refining* (US Dept. Energy, 2015).
4. Kim, J. *et al. Sep. Sci. Technol.* **48**, 367–387 (2013).
5. Koros W. J. & Lively, R. P. *AIChE J.* **58**, 2624–2633 (2012).
6. Xu, L. *et al. J. Membr. Sci.* **423–424**, 314–323 (2012).
7. Interagency Working Group on Social Cost of Carbon (US Govt.). *Social Cost of Carbon for Regulatory Impact Analysis* (2013).
8. Song, C. *Catal. Today* **115**, 2–32 (2006).
9. Jordens, A., Cheng, Y. P. & Waters, K. E. *Miner. Eng.* **41**, 97–114 (2013).
10. Massari, S. & Ruberti, M. *Resour. Policy* **38**, 36–43 (2013).

CORRECTION

The graphic 'The dirty ten' in the Comment 'Three steps to a green shipping industry' (Z. Wan *et al. Nature* **530**, 275–277; 2016) gave the wrong unit for PM_{2.5} concentrations. It should have been $\mu\text{g per m}^3$, not mg per m^3 .

Beyond the 'InterNyet'

Michael D. Gordin reviews a history of the Soviets' failed national computer network.

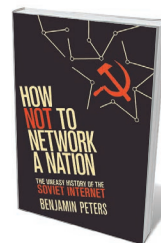


Soviet computer scientist Victor Glushkov pushed to create an 'all-union' system.

The 'small- n ' challenge plagues all historians, but this problem of undersized samples is especially acute in the history of science and technology. Most scientific discoveries seem to happen uniquely. We do sometimes see multiples — about half a dozen articulations of the principle of the conservation of energy or the periodic table around the same time, for example — but the diversity of specializations, the pace of communication and the vagaries of publishing mean that most innovations arise as singletons, to use sociologist Robert Merton's term. The issue is perhaps most striking with the

flagship technology of our present moment: the Internet. Here, we have an n of 1.

This matters for two reasons. First, singletons frustrate generalization, making it difficult to draw lessons for science policy. Second is the related puzzle of contingency. We currently have an Internet, and it has a set of properties (such as the end-to-end principle, which stipulates that applications should happen at the edges of the network, rather than at intermediary nodes). Does it look like this because it has to, or are its features contingent characteristics of this specific Internet? Without alternatives to compare it



How Not to Network a Nation: The Uneasy History of the Soviet Internet
BENJAMIN PETERS
The MIT Press: 2016.

to — a larger n — we just cannot say.

In *How Not to Network a Nation*, communications specialist Benjamin Peters argues for contingency, on the basis of an n increased from 1 to 2. Well, to 1.37 or thereabouts.

Historians have already started to chronicle networks past as useful comparisons. One is Project Cybersyn, an experiment to network the Chilean economy under president Salvador Allende in the 1970s, described in Eden Medina's *Cybernetic Revolutionaries* (MIT Press, 2011). Peters summarizes these well, but his quarry is the great white whale of this specialized historiography: the Soviet Internet. Whether there ever was such a thing, why it never moved beyond the project stage and which of the various projects between the late 1950s and the late 1980s can be properly classified as efforts to develop one are the main subjects of the book.

Peters makes a good case to move beyond historian Slava Gerovitch's excellent pun on this seeming oxymoron: "InterNyet". His intuition is spot on. The cold-war origins of the US networking programme have been well documented, for example in Janet Abbate's *Inventing the Internet* (MIT Press, 1999). Direct military sponsorship was crucial. The defence department provided patronage through its Advanced Research Projects Agency, which launched ARPANET, the embryonic Internet, on 29 October 1969. The Internet's conceptual roots included cybernetics, created by mathematician Norbert Wiener in 1948. Given the close parallels between Soviet and US cold-war technologies, it would be surprising if there had been no efforts to generate a Soviet Internet. Indeed, Peters finds six different proposals to develop an 'all-union' computer network. This stands to reason, given what he calls "the outsized infrastructural imagination of Soviet planners", who liked their projects big and utopian — think the space programme, dams and nuclear power.

Peters concentrates on computer scientist Viktor Glushkov's OGAS (*obshchegosudarstvennaia avtomatizirovannaia*

SPUTNIK/ALAMY

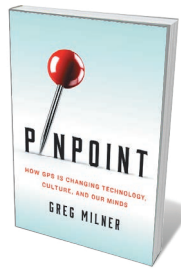
sistema, as the standard Library of Congress transliteration would render it — although various systems are used inconsistently throughout the book). The full name is a mouthful: “All-State Automated System for the Gathering and Processing of Information for the Accounting, Planning, and Governance of the National Economy, USSR”. Beginning in 1962, Glushkov spent 25 years trying to mobilize support for his network from his Institute of Cybernetics in Kiev, which created a rich set of cultural resources, including a model constitution, passport and cartoons depicting the land of “Cybertonia”. Peters reproduces these in plentiful images and descriptions, chronicling their utopian spirit and demonstrating the need for engineers in all times to let off steam through flights of fancy. But the project was never realized.

It is difficult to glean all the technical specifics from the material that Peters mobilizes from archives, interviews and declassified CIA reports. Some proposals look like cloud computing or tablets, but it would be anachronistic to interpret them in that way (and Peters doesn’t). The idea was to use real-time processing to connect economic inputs and outputs, rendering the planned economy both functional and adaptive. We cannot even be sure that Glushkov’s plans would have worked. What we do know is that the failure was not caused by a scarcity of personal computers, because OGAS was meant to link factory mainframes. Nor was it ideology: cybernetics, as Peters readably recounts, was well suited to Soviet ideological preferences in materialism and planning. To discover the roots of the issue, Peters invokes the cybernetic concept of heterarchy, which he defines as “complex networks with multiple conflicting regimes of evaluation in operation at the same time”; he then uses this to explore the heterogeneity of approaches to networking.

Perhaps predictably, OGAS’s demise was death by a thousand paper cuts. Documents were misfiled, meetings were missed, the military and the statistical ministries disagreed about who would benefit. Peters’s provocative thesis is that “The capitalists behaved like socialists while the socialists behaved like capitalists.” The US Internet was the result of state subsidies and benevolent paternalism; the Soviet attempt foundered on entrepreneurial infighting. (Elsewhere, Peters puts the culprit down as cost, although how costs were tabulated was in itself a bureaucratic conundrum.) There is no dramatic climax to *How Not to Network a Nation*. Non-existent technologies end with a whimper, but even whimpers can tell you something. ■

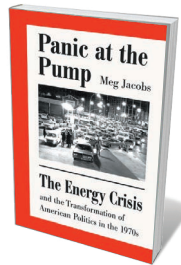
Michael D. Gordin is Rosengarten Professor of Modern and Contemporary History at Princeton University in New Jersey. His most recent book is *Scientific Babel*. e-mail: mgordin@princeton.edu

Books in brief



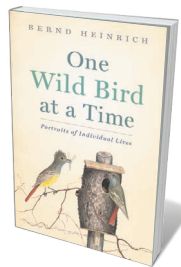
Pinpoint: How GPS is Changing Technology, Culture, and Our Minds Greg Milner W. W. NORTON (2016)

It is key to the Internet’s operation, is deployed in seismology and climate-change research — and can lead drivers into seriously tight spots. The multisatellite Global Positioning System (GPS), reveals journalist Greg Milner in this assured technological history, is a risk-laden ubiquity that has profoundly changed society. He traces its conceptual and practical roots from early Polynesian navigational acumen through cold-war US prototypes to today’s system, kick-started by Bradford Parkinson. Milner delves, too, into the cognitive impacts of reliance on GPS, and ethical issues around data misuse.



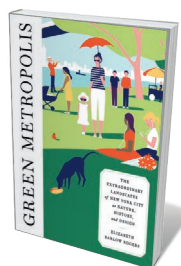
Panic at the Pump: The Energy Crisis and the Transformation of American Politics in the 1970s Meg Jacobs HILL AND WANG (2016)

In 1973, ‘oil shock’ engulfed the United States as the Organization of Arab Petroleum Exporting Countries embargoed exports. Historian Meg Jacobs incisively chronicles the ensuing policy war, as the Nixon administration and free-marketeers called for deregulation of the market, and the left (including Henry ‘Scoop’ Jackson, Democratic senator for Washington) pushed for alternative energy. That battle, Jacobs argues, reverberates in fracking and climate-change policy today, and offers lessons for the transition to a fossil-free future.



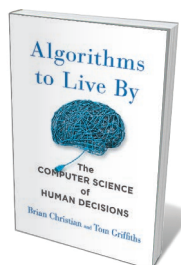
One Wild Bird at a Time: Portraits of Individual Lives Bernd Heinrich HOUGHTON MIFFLIN HARCOURT (2016)

Biologist Bernd Heinrich’s cabin in the Maine woods is a “live-in bird blind” engineered for year-round observation, and his engrossing scientific memoir lets us in on the ornithological action. Here are northern yellow-shafted flickers nesting in a wall cavity (Heinrich estimates it takes 21,600 ants to fledge one nestling); an avian soundtrack veering from the cackling of a barred owl to the “tinkles, whistles, twitters, growls, and squawks” of a common starling; and a woodcock bursting into rocket-like flight. Step by finely attuned step, we learn, with Heinrich, “one wild bird at a time”.



Green Metropolis: The Extraordinary Landscapes of New York City as Nature, History, and Design Elizabeth Barlow Rogers KNOPF (2016)

New York may seem the archetypal cityscape, but nature thrums through this concrete jungle. So reports landscape preservationist Elizabeth Barlow Rogers in her erudite study of seven of the city’s green spaces. Summoning geology, biology and history, Barlow witnesses stridulating 17-year cicadas at Staten Island’s High Rock Nature Center, walks through the 14.5-hectare “self-generating wildwood” of Central Park’s Ramble, strolls the evocative garden promenade of reclaimed rail spur the High Line, and more.



Algorithms to Live By: The Computer Science of Human Decisions Brian Christian and Tom Griffiths HENRY HOLT (2016)

When do you cut short a house search? How do you schedule a day’s worth of tasks? Meshing psychology with computational models, writer Brian Christian and cognitive scientist Tom Griffiths argue that algorithms are ace tools for solving the pressing conundrums that litter life. Far from being narrowly prescriptive, their algorithmic fixes (such as the 37% rule, otherwise known as the secretary problem) are forgiving — not least, in showing how messiness can sometimes be an optimal choice. [Barbara Kiser](#)

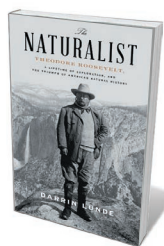
In the museum with Roosevelt

Michael Ross Canfield enjoys a chronicle of the statesman's natural-history legacy.

The head of a Cape buffalo presents itself just inside the door of Theodore Roosevelt's historical home, Sagamore Hill, on Long Island, New York. A few steps further in are mounted rhinoceros horns, then a trophy room framed by elephant tusks. This is, in effect, the personal natural-history museum of the explorer, soldier and 26th US president. Roosevelt also donated hundreds of specimens to the American Museum of Natural History in New York and the Smithsonian Institution in Washington DC. Between these two kinds of museum — the private and the public — we find the Roosevelt of *The Naturalist* by Darrin Lunde, manager of the Smithsonian's mammal collections.

Lunde's narrative stretches from Roosevelt's youth to his return from a scientific safari in what is now Kenya in 1909–10, a decade before his death. Roosevelt collected and preserved specimens throughout his life. He chronicled his hunts (along with bar fights and chasing outlaws) in popular books such as *Hunting Trips of a Ranchman* (1885), and continued to collect everything from manta rays in Florida to moose in Canada through his presidency (1901–09) and after. As Lunde reveals, his bursts of field work coincided with — and fed into — the evolving US scientific study of nature that is fostered by museum founders, such as Albert Bickmore of the American Museum of Natural History. Roosevelt's activities filled museums and inspired him to use his political mandate to protect 93 million hectares of public land and establish 5 national parks.

Born in 1858 into a wealthy Manhattan household, the home-schooled Roosevelt avidly read natural histories in the family library and ferreted out animals in the wilder surrounds of New York City. A dead seal that he encountered in a Broadway market when he was around seven made a singular impression; he recorded measurements of it and acquired the head. He established a collection ('The Roosevelt Museum of Natural History') in his bedroom, and a natural-history society with his peers — complete with



The Naturalist: Theodore Roosevelt, A Lifetime of Exploration, and the Triumph of American Natural History
DARRIN LUNDE
Crown: 2016.



Theodore Roosevelt, 26th US president.

curatorial specialities such as conchologist, and papers on topics such as the migration of whales. In his early teens, he studied taxidermy with John Bell, who had worked with naturalist-illustrator John James Audubon. Later, at Harvard University in Cambridge, Massachusetts, he began to study biology, but eventually gravitated toward economics and history — key preparation for a dual career in statesmanship and conservation.

Roosevelt's field work, like that of most museum naturalists at the time, revolved around specimen collection and preparation, and the window that Lunde opens on this is among the book's novel contributions. Taxidermy involved applying arsenical soap to skins, boiling bones and allowing bacteria or beetles to eat flesh. It was hard and unpleasant, and on Roosevelt's African trip, professionals such as Edmund Heller did much of it for him: Heller "roughed out" specimens by carving soft tissue from the hides and bones. Even as a revered ex-US president in Africa, however, Roosevelt never shied away from close observation of specimens. Taxidermist Carl Akeley photographed him holding a camera while investigating an elephant carcass with a hyena scavenging inside it.

Roosevelt's youthful collecting technique was basic, and included knocking birds' nests from trees; he switched to guns in maturity. Like many hunter-naturalists up to the late twentieth century, he both loved and

killed animals. This contradiction has exercised many. Teasing apart aspects of ethics, morality, manliness and environmentalism in Roosevelt's approach to collecting, Lunde reveals how the president's impulses overlapped. He hunted for meat and sport — a common pursuit among the wealthy on both sides of the Atlantic — as well as science.

That scientific strand was strong. Lunde describes how Roosevelt was able to "hold specimens in his hand", whether bear, cougar or bird, to hone his observational acuity. Roosevelt even chastised hunters who did not learn in this way and report results appropriately, because information could easily be lost to science. Other areas of his life, particularly his approach to politics and policymaking, show the imprint of these habits of observing, collecting disparate elements and information, and analysing assembled parts.

Yet Roosevelt spent relatively little time actually in museums. These were then growing into prominence under pioneers such as Spencer Fullerton Baird, the first curator at the Smithsonian, and C. Hart Merriam, who expanded the scientific study of animals at the US Department of Agriculture, both of whom Lunde discusses. Professional curatorship was not for Roosevelt. Even at Harvard he gravitated toward field work, largely dismissing the focus on section cutting and minutiae taught at Louis Agassiz's museum there. Lunde's remark that exploring museums is "like traveling around the world" reflects Agassiz's view that assemblages of specimens allow a naturalist to read from "the great book of nature". But Roosevelt was not satisfied with simply reading. He wanted to write his own accounts of the wild.

As a curator, Lunde might have shared more about the scope of Roosevelt's collections and their current value, particularly in an age of unprecedented biodiversity loss. However, *The Naturalist* does highlight the crucial importance of maintaining such legacies. It also helps to disentangle Roosevelt's roles as hunter, conservationist and museum man — and for anyone visiting Sagamore Hill, it enriches contemplation of objects such as the bearskin rug or rhino-foot inkwell. ■

Michael Ross Canfield lectures on organismic and evolutionary biology at Harvard University in Cambridge, Massachusetts, and studies how scientists record their work. His latest book is *Theodore Roosevelt in the Field*. e-mail: canfield@fas.harvard.edu

Correspondence

Benefits of trade in amber fossils

Amber of great palaeontological significance is flowing into China's jewellery market, fuelling a trade that dates back some 13,000 years. Ironically, banning this trade could be more damaging to science than letting it continue.

Fossiliferous ambers are being extensively destroyed by mining activity. The renowned Zhangpu amber from southeast China, for example, is being burned in the process of kaolin extraction. The Fushun amber site is closing after more than 110 years of adjacent lignite mining (B. Wang *et al. Curr. Biol.* **24**, 1606–1610; 2014).

Amber affords exceptional preservation of insects and microorganisms, shedding light on ephemeral behaviours such as parasitism, predation and camouflage. These fossils often provide more detail than rock fossils about an organism's morphology, ecology, ethology and evolutionary history (see, for example, D.-Y. Huang *et al. Sci. Rep.* **6**, 23004; 2016).

Amber excavation involves manpower and materials that are not available to palaeontologists. The jewellery trade instead provides them with the organismal inclusions, either directly as unwanted material or indirectly by preserving the fossils in finished gems for posterity.

Jun Chen *Linyi University, China.*

Bo Wang *Nanjing Institute of Geology and Palaeontology, China.*

Edmund A. Jarzembowski *Natural History Museum, London, UK.*

rubiscada@sina.com

Postdoc rights need not hurt productivity

Portugal's government is on the verge of a historic process, recognizing at last that postdoctoral researchers should have the same rights as the rest of the country's workforce

(see go.nature.com/famkkn; in Portuguese). In defiance of European Union practice, more than 90% of these early-career scientists are currently classed as 'advanced students' and funded by student scholarships.

Some institutions are expected to resist this change. The idea persists that big research teams boost scientific productivity, with postdocs offering the cheapest means of expansion.

Changing the scholarship system to non-tenured work contracts — comparable to postdoc fellowships in most developed countries — will mean losing roughly one-third of postdocs, assuming government funding stays the same. Even so, 96% of postdocs surveyed at the University of Minho support the change. A much bigger survey by the Portuguese National Association of Researchers in Science and Technology is expected to yield similar results.

International evidence indicates that reducing the number of postdocs by one-third is unlikely to impair productivity (see *Nature* **531**, 263–265; 2016). **Nuno Cerca** *University of Minho, Braga, Portugal.* *nunocerca@ceb.uminho.pt*

Ventilating Beijing cannot fix pollution

Beijing plans to build a system of ventilation corridors across the city to help dissipate heat and smog (see go.nature.com/cgbd7i). We suggest that a more comprehensive solution is needed to tackle the scale and complexity of Beijing's severe air pollution.

In our view, the city's situation in a valley ringed by mountains — combined with the fall in average winter wind speed over the past decades (Z. Li *et al. Adv. Atmos. Sci.* **28**, 408–420; 2011) — is likely to reduce the effectiveness of these corridors, particularly in winter, when smog is most severe. The prevailing northwesterly winter wind could also propel

dangerous particulate matter and other pollutants along the corridors to the south of the city, where the wind gradually weakens and so is less effective at dispersal (X. He *et al. Build. Environ.* **92**, 668–678; 2015).

Measures to control air pollution need to address causes as well as symptoms. Industrial structures, energy inefficiency, the pursuit of economic growth and the extent of regional cooperation all contribute (Y. Liu *et al. Nature* **517**, 145; 2015). These must eventually be brought into line through integrated, scientifically informed planning.

Yansui Liu *Beijing Normal University, China.*

Yang Zhou, Yurui Li *Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China.*

liuys@igsnrr.ac.cn

Anti-science wave sweeps Poland

Poland's government is showing a worrying trend to disregard scientific evidence and rationality (see, for example, *Nature* **530**, 393; 2016). Polish academia needs the backing of international scientific societies to help counter some alarming implications for the population.

For instance, we find it questionable that Poland's Parliamentary Committee on the Safety of the Programme of Vaccination of Children and Adults invited an anti-vaccination activist to speak as an expert at one of its meetings (see go.nature.com/6bbg5u; in Polish).

In addition, the country's *in vitro* fertilization (IVF) programme has been axed. This leaves some 17,000 couples stranded in mid-treatment and almost 22,000 others on the waiting list (see go.nature.com/nfuede; in Polish). And if proposed anti-abortion legislation goes through, IVF is likely to become illegal. The

health ministry is developing an alternative programme based on natural procreative technology ('naprotechnology'), a fertility treatment that is approved by the Catholic Church but lacks sound scientific support.

Some Polish universities, such as the University of Gdańsk and the Medical University of Wrocław, are engaging speakers on such scientifically refuted topics as curing cancers with vitamin C or breast enlargement through hypnosis. Creationism, too, seems to be experiencing a resurgence. For example, the book *Ewolucja, Dewolucja, Nauka* (*Evolution, Devolution, Science*) (Frona, 2016) by the dendrologist Maciej Giertych, which we read as arguing against evolution, is being promoted in schools. In our view, this poses a threat to the country's scientific-education programme.

Paula Dobosz *University of Cambridge, UK.*

Jakub Zawila-Niedzwiecki *University of Warsaw, Poland.* *jakub@zawila-niedzwiecki.pl*

Tell us the end of the story

Nature's Correspondence section is known for highlighting political aspects of scientific issues of public interest and for calling on organizations and governments to address scientists' concerns in developing policy. Readers would surely like to know what happened next.

Contacting the authors could be enlightening, but readers' curiosity may not transcend their busy schedules. Perhaps correspondents should consider adding an online footnote to their letter to indicate its impact.

M. Usman *University of Agriculture, Faisalabad, Pakistan.*

A. Chaudhary *Government College University, Faisalabad, Pakistan.*

M. Farooq *College of Veterinary and Animal Sciences, Jhang, Pakistan.* *muhammad.usman@uaf.edu.pk*

R. McNeill Alexander

(1934–2016)

Zoologist who pioneered comparative animal biomechanics.

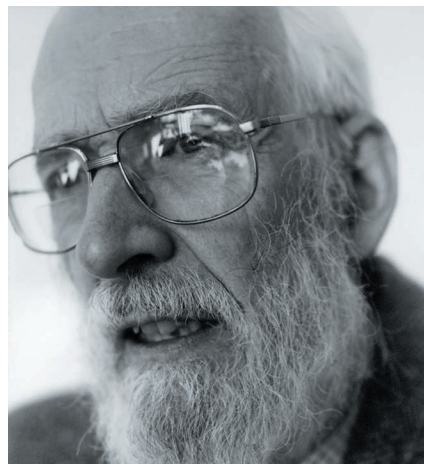
Robert McNeill Alexander defined many fundamental properties of how animals move. He combined elegant mechanical and mathematical analyses to reduce problems in flight, swimming, walking, running and anatomy to their simplest level. He explained the importance of inertial forces versus gravitational ones in determining which gaits land animals use to move at different speeds, and he predicted the pace at which dinosaurs probably moved.

His work showed why animals of different sizes move in similar ways, as well as the importance of the elastic energy stored in tendons in reducing the metabolic cost of jumping and running. Alexander wrote 20 books, including his classic text *Animal Mechanics* (Sidgwick and Jackson, 1968). He published more than 280 scientific papers. On most subjects in biomechanics, it would be wise to first read what Alexander had to say.

Born in 1934 in Lisburn, UK, to the chief engineer of the city of Belfast Robert Alexander and author Janet McNeill, he was inspired at school by his biology teacher Arnold Benington, a BBC radio naturalist. Aged 16 or 17, Alexander published his first paper 'Behaviour of the robin during laying' (*Brit. Birds* **44**, 389–390; 1951), after a pair nested on top of a wardrobe in his bedroom. That year, he also won an essay prize from the Royal Society for the Protection of Birds for an experiment to test if birds would remember which trough on his windowsill contained hidden food.

At the University of Cambridge, Alexander read natural sciences. He completed a PhD on the function of swim bladders in fish. His supervisor was James Gray, a pioneer of comparative experimental studies of animal locomotion. He also acted, including a memorable appearance as the giant Harapha in a production of John Milton's *Samson Agonistes*. And he travelled extensively in Europe, meeting his wife Ann, also a student at Cambridge, on a trip to Italy in 1956. He was an expedition scientist on a Cambridge trip to the jungles of Guyana in 1960. After a lectureship at the University College of North Wales (now Bangor University) from 1958 to 1969, Alexander became professor of zoology at the University of Leeds, until his retirement in 1999.

In the early 1960s, the functional analysis of animal form was driven by descriptive comparisons of morphology that largely lacked mathematical expression. Alexander addressed this by combining field studies with laboratory experimentation and



theoretical modelling. In collaboration with Harvard University physiologist C. Richard Taylor and Kenyan veterinary scientist Geoffrey Maloiy, he examined movement in African mammals as diverse as dik-dik and buffalo. He formulated a model that explained how and why animals move in similar ways, and he used Froude numbers (previously deployed by Victorian engineer William Froude in his analysis of the bow waves of ships) to explain how the length of limbs affects speed and gait over ground. Alexander also built theoretical models of foraging and migration, suggesting that only birds or large mammals benefit from the risks and energy costs of long-distance travel.

Alexander compared the athletic performance of humans with that of other animals. He discovered that small animals rely on the rapid release of elastic energy stored in their tendons to jump high and far, whereas humans and larger animals stretch their muscles with force to achieve greater heights. And he evaluated the evolutionary and energetic consequences of building and maintaining physiological and mechanical structures that are necessary for locomotion, such as those that facilitate respiratory gas transport and musculoskeletal support.

His family indulged his research endeavours. In the 1970s, while developing an approach to derive the speeds of dinosaurs from their fossilized tracks, Alexander took his two children to Snettisham beach in Norfolk, UK. There, they walked and ran along various textures of mud, counting their strides and timing themselves with a stopwatch. These antics are recorded in the paper 'Estimates of speeds of

dinosaurs' (*Nature* **261**, 129–130; 1976).

Alexander's work fuelled the emerging field of biorobotics. He participated in a European effort to build a robot dinosaur, advising on the probable gait and a simplified arrangement of joints for the creature. His findings also contributed to improved gait rehabilitation and prosthetic devices for people. His coffee-table book *Bones* (Prentice Hall, 1994) reveals the beauty inherent in the biomechanics of animals. His 1995 educational CD *How Animals Move* was widely used in schools and universities; it remains the best teaching aid of its kind.

Neill was devoted to comparative biomechanics and its wider appreciation. He served as secretary to the Zoological Society of London from 1992 to 1999 (after the reversed decision to close London Zoo). He served as president of the Society for Experimental Biology and the International Society of Vertebrate Morphology and as editor of the journal *Proceedings of the Royal Society B*. Among his many honours, Neill was elected fellow of the Royal Society in 1987 and appointed Commander of the Order of the British Empire in 2000. The gong that gave him the most amusement — and best captured his Gandalf-like status — came in 1996, when British newspaper *The Mail on Sunday* listed him as one of "Britain's Nuttiest Professors — Ten Sages Who Really Know Their Onions".

In retirement, Neill continued to attend professional meetings, give lectures and serve on examination committees. He advised on television documentaries, including the BBC's 2001 series *Walking with Beasts*. As a reviewer, he was always succinct, insightful and supportive of good science. He visited every poster at conferences, where he conveyed his passion to students and remained a role model. Neill was warm and animated in conversation and broad-minded when communicating his science. He was an inspiration and generous mentor to many of today's leaders in the scientific field that he established. ■

Andrew A. Biewener is professor of biology at Harvard University in Cambridge, Massachusetts, USA, and director of the Concord Field Station, where he collaborated with Neill Alexander. **Alan Wilson** is professor of locomotor biomechanics at the Royal Veterinary College in London. Alexander examined Wilson's PhD thesis. e-mails: biewener@fas.harvard.edu; awilson@rvc.ac.uk.

JOHN ARNISON/NATIONAL PORTRAIT GALLERY, LONDON

OCEAN SCIENCE

The rise of Rhizaria

Large amoeba-like organisms known as Rhizaria have often been overlooked in studies of ocean biology and biogeochemistry. Underwater imaging and ecological network analyses are revealing their roles. [SEE ARTICLE P.465](#) & [LETTER P.504](#)

DAVID A. CARON

Do you know the name and evolutionary affiliation of any of the most conspicuous groups of single-celled organisms in the world's oceans? Did you guess the Rhizaria, or one of the more familiar groups of plankton that make up this supergroup, such as the Radiolaria, Acantharia or Foraminifera? If you didn't, you're not alone — until recently, neither did the vast majority of biological oceanographers. Biard *et al.*¹ report on page 504 of this issue that the abundance and biomass of these enigmatic species in the ocean are much greater than previously recognized. In addition, Guidi *et al.*² (page 465) reveal the extent of the Rhizaria's involvement in the export of carbon from the atmosphere to the ocean depths.

Oceanic Rhizaria are protists: single-celled and some colonial organisms that are eukaryotic, meaning that they contain nuclei and other membrane-bound organelles. The Rhizaria were formerly thought to be phylogenetically related to the much smaller and better known amoebae, because both groups feed by capturing and engulfing prey with extensions of their cytoplasm called pseudopodia. However, the Rhizaria can produce complex pseudopodial networks that attain sizes of more than a centimetre. Some species can even form cylindrical colonies approximately 1 cm in diameter and greater than 1 m in length³.

These pseudopodial networks, and the intricate mineral skeletal structures of opal (SiO₂), celestite (SrSO₄) or calcite (CaCO₃) that many Rhizaria form, distinguish them from amoebae, as does DNA-sequence information. The supergroup Rhizaria was devised

more than a decade ago to contain these morphologically complex forms, and their smaller amoebic cousins have been placed among several eukaryotic supergroups in modern phylogenetic schemes⁴.

detected by divers in the open ocean more than two decades ago^{9,10}, and are visible in earlier underwater images¹¹. However, truly global surveys have never been conducted.

The *Tara* Oceans project has begun to

The large oceanic Rhizaria entangle and engulf a wide range of prey in their pseudopodial networks⁵. Many species dwelling in the upper ocean also possess symbiotic algae⁶, which can contribute significantly to host nutrition and to total primary production in the ocean⁷. This nutritional versatility makes amoeboid Rhizaria well adapted for life in the vast stretches of oligotrophic (nutrient-poor) waters of the open ocean.

The renowned nineteenth-century German scientist and artist Ernst Haeckel immortalized these species in drawings that captured their elegance and complexity (Fig. 1). Much of the material for Haeckel's drawings came from samples returned by the *Challenger* expedition of 1872–76, a circumnavigation of the planet that laid the foundation for modern oceanography⁸. Yet, although the Rhizaria are valued by palaeontologists for climate reconstructions based on the fossil shell assemblages left by some of these species in deep ocean sediments, they have received only scant attention from biologists.

One of the reasons for their anonymity to oceanographers is the delicate morphologies of living specimens. These structures deteriorate badly as a result of the methods and preservatives that have routinely been used for collection and species identification. Some species contain no skeletal material, and in plankton samples their remains are often not recognizable. Substantial abundances of Rhizaria were

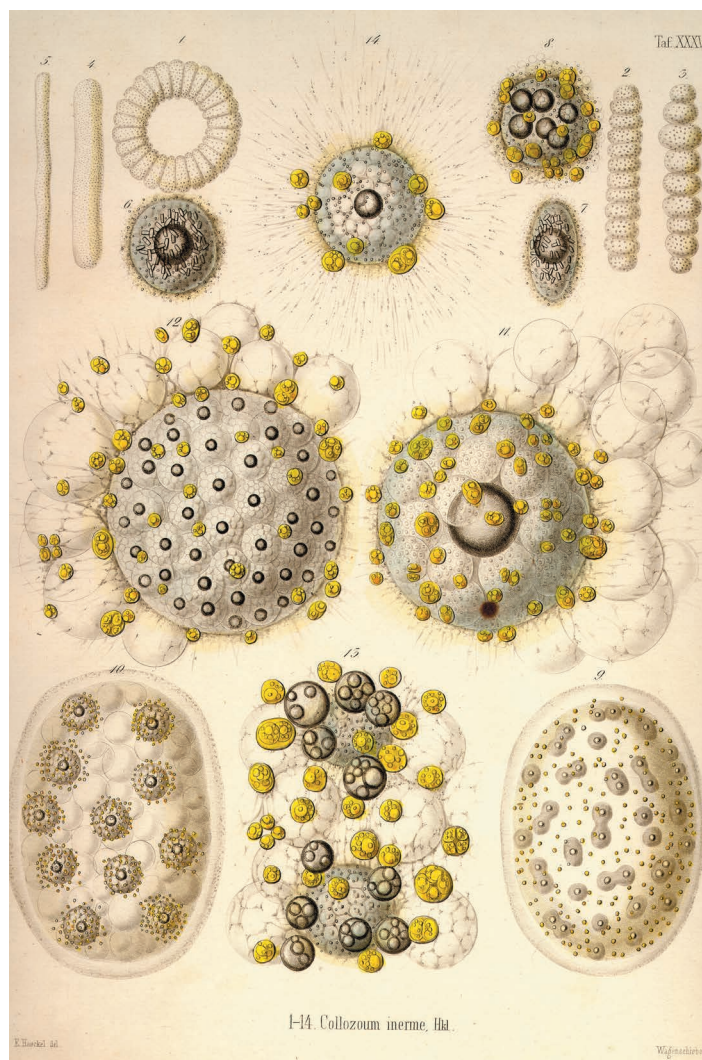


Figure 1 | Abundant plankton. These illustrations by Ernst Haeckel were drawn from samples collected by the oceanic *Challenger* expedition of 1872–76. They depict the colony shape, central capsule structure and symbiotic algae in the colonial plankton *Collozoum inerme*, which belongs to the supergroup Rhizaria. Biard *et al.*¹ and Guidi *et al.*² present analysis of data collected from the *Tara* Oceans expedition that reveals an unexpectedly large abundance of Rhizaria in the ocean, and implicates these organisms in the vital export of carbon from upper ocean layers to the deep ocean.

BUYENLARGE/GETTY

address this gap. The *Tara* schooner has circumnavigated the globe, conducting extensive sampling of the biological communities and surveying the environmental conditions in the upper layer of the ocean, with the goal of enhancing our understanding of its organismal and genetic biodiversity, and the biogeochemical cycles affected by these communities. The data include a variety of oceanographic measurements (such as temperature, salinity and light), as well as the size distributions of plankton and estimates of sinking-particle flux. Biological data obtained include vast numbers of underwater images, genetic 'bar-coding' of all plankton — ranging from viruses to multicellular zooplankton — and combined genomic (metagenomic) data for bacteria, archaea, viruses and minute eukaryotes.

Biard *et al.* analysed nearly 2 million of the underwater images collected during the expedition¹², and concluded that abundances of large Rhizaria in the global ocean have been greatly underestimated by conventional sampling methods. On the basis of abundance and volume, the authors estimate the collective carbon content of these species to be around 10^{14} grams of carbon in the upper 200 m of the ocean. If accurate, this biomass places oceanic Rhizaria on a par with other large, 'conventional' zooplankton groups in the ocean, such as krill.

Guidi *et al.* used regression-based modelling and weighted gene-correlation network analysis to determine correlations between the project's genetic information and the sinking of carbon-containing particles. Sinking particles are an important component of the ocean's biological carbon pump — a mechanism by which carbon is removed from surface waters for periods of up to tens of thousands of years, thus helping to reduce atmospheric carbon dioxide concentrations. The analysis revealed some expected relationships, such as a correlation between carbon flux and the presence of small crustaceans called copepods, which produce rapidly sinking faecal pellets. Among the surprising results, however, is Guidi and colleagues' implication of large Rhizaria (specifically, the genetic bar codes of several radiolarian groups) as key players in the export of material from the upper ocean.

This finding makes sense, in that large plankton are thought to have a disproportionately greater role in particle flux than small particles and organisms, and because many Rhizaria form dense crystalline structures, which may increase sinking rates. Moreover, the findings are consistent with the high abundances of Rhizaria established by Biard and colleagues.

Considered together, the two studies provide the first quantitative assessments of the role of large Rhizaria in the ocean: the organisms' abundance, biomass and relationship to sinking particles. Much additional work will be needed to fully characterize the vertical,

geographical and seasonal distributions of these species, and how they might respond to changing climatic and oceanic conditions. For example, Biard *et al.* speculate that global abundances of Rhizaria may increase if oligotrophic oceanic realms expand, as predicted in some climate-change scenarios.

For the moment, the studies have created awareness of the global significance of large Rhizaria, and provided evidence of the insufficiency of conventional sampling methods for estimating their abundances. This work is a fitting sequel to Haeckel's seminal work on these beautiful creatures, albeit more than a century later. ■

David A. Caron is in the Department of Biological Sciences, University of Southern California, Los Angeles, California 90089-0371, USA.
e-mail: dcaron@usc.edu

MATERIALS SCIENCE

Cracks help membranes to stay hydrated

Membranes have been prepared with a cracked coating that prevents them from drying out in low-humidity conditions — a boon for devices, such as fuel cells, that need hydrated membranes to function. SEE LETTER P.480

JOVAN KAMCEV & BENNY D. FREEMAN

The synthetic polymer membranes used in fuel cells, water purifiers and systems for harvesting electricity from the sea must be hydrated — desiccation diminishes their performance. This is a problem, because some of these applications (including fuel cells) operate at high temperatures in low-humidity environments. On page 480 of this issue, Park *et al.*¹ describe membranes that limit their own dehydration, substantially improving the membranes' performance in low-humidity environments.

The membranes in question are called ion-exchange membranes (IEMs), and they are made from polymers in which acidic or basic chemical groups are covalently bound to a polymer backbone². IEMs swell when in contact with water, causing the attached groups to dissociate into fixed and mobile ions, and so making the membranes highly charged. The membranes therefore selectively and efficiently transport counter-ions (ions that have an opposite charge to that of the polymer's charged groups), so that rates of ion transport through the membranes are high. This high ionic conductivity is crucial for many membrane-based technologies because it reduces

1. Biard, T. *et al.* *Nature* **532**, 504–507 (2016).
2. Guidi, L. *et al.* *Nature* **532**, 465–470 (2016).
3. Anderson, O. R. *Radiolaria* (Springer, 1983).
4. Pawłowski, J. & Burki, F. *J. Euk. Microbiol.* **56**, 16–25 (2009).
5. Swanberg, N. R. & Caron, D. A. *J. Plankton Res.* **13**, 287–312 (1991).
6. Decelle, J., Colin, S. & Forster, R. A. in *Marine Protists: Diversity and Dynamics* (eds Ohtsuka, S., Suzuki, T., Horiguchi, T., Suzuki, N. & Not, F.) 465–500 (Springer, 2015).
7. Caron, D. A., Michaels, A. F., Swanberg, N. R. & Howse, F. A. *J. Plankton Res.* **17**, 103–129 (1995).
8. Haeckel, E. in *Report on the Scientific Results of the Voyage of H.M.S. Challenger During the Years 1873–1876 Vol. 18* (ed. Thompson, C. W.) 1–1803 (HMSO, 1887).
9. Swanberg, N. R. *Limnol. Oceanogr.* **28**, 655–666 (1983).
10. Michaels, A. F., Caron, D. A., Swanberg, N. R., Howse, F. A. & Michaels, C. M. *J. Plankton Res.* **17**, 131–163 (1995).
11. Dennett, M. R., Caron, D. A., Michaels, A. F., Gallager, S. M. & Davis, C. S. *J. Plankton Res.* **24**, 797–805 (2002).
12. Karsenti, E. *et al.* *PLoS Biol.* **9**, e1001177 (2011).

This article was published online on 20 April 2016.

energy losses and therefore lowers costs.

Water in IEMs facilitates ion transport by keeping ions hydrated and ionized, and by creating or enlarging gaps between polymer chains to reduce the ability of the bulky, slow-moving chains to impede transport of the much smaller, highly mobile ions. In other words, ionic conductivity depends sensitively on water content in IEMs. The membrane's water content is closely linked to membrane swelling — membranes swell as they absorb water and shrink as they lose it. When humidity is low and temperatures are high (both of which are advantageous for applications such as fuel cells), IEMs become desiccated, swelling is reduced relative to that at high humidity and ionic conductivity falls. Controlling IEM dehydration under such conditions has been a formidable technical and scientific barrier that has limited membrane performance.

Park *et al.* have developed a strategy that helps membranes to retain water in low-humidity environments. They deposited a thin layer (8–260 nanometres thick) of a highly water-repellent material on the surface of IEMs (approximately 50 μm thick) in a relatively low-humidity environment (40% relative humidity), thus endowing the hydrophilic, but partially dehydrated,

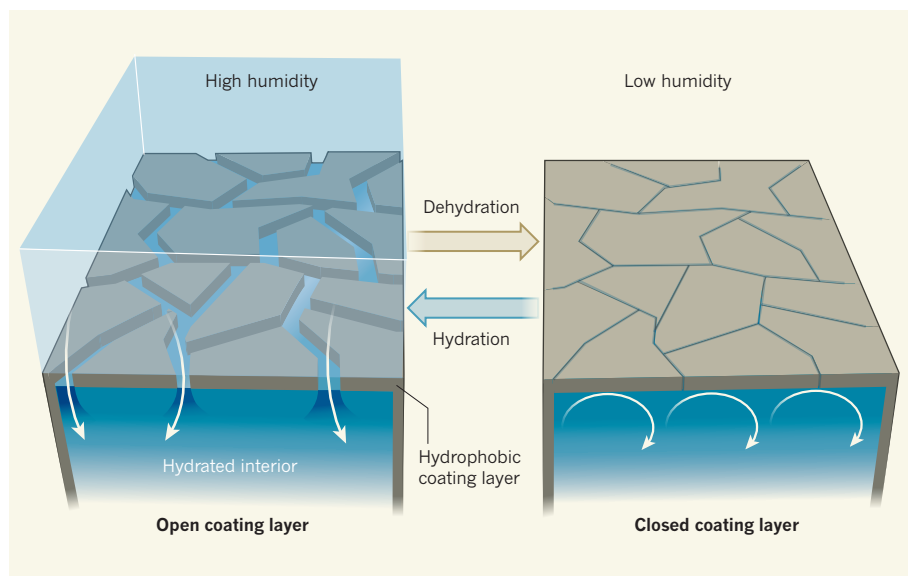


Figure 1 | An ion-exchange membrane that responds to humidity. Ion-exchange membranes allow ions to pass through them, and are needed for applications such as fuel cells. Some water must be present in the membrane to support ion transport, but water is often lost from membranes in conditions of low humidity and/or high temperature. Park *et al.*¹ have prepared membranes coated with a thin layer of a hydrophobic material that contains cracks. At high humidity, the membrane swells as it absorbs water (white arrows indicate water movement), mainly through the cracks, allowing ion transport. At low humidity, the membrane loses some water and contracts. This substantially closes the cracks and slows water loss, helping the membrane to retain moisture and maintaining efficient ion transport even in dry conditions.

membranes with a hydrophobic skin.

Such a skin would typically be almost impervious to both ions and water, and would therefore sharply reduce the membrane's ionic conductivity. But when the authors exposed their membranes to high humidity, some water migrated through the thin hydrophobic coating into the highly hydrophilic interior, causing the membrane to swell sufficiently to generate 36–340-nm-wide cracks in the coating. These nanocracks expand in high-humidity environments (100% relative humidity) as the membrane swells, allowing the membrane to absorb more water and rapidly transport ions, but contract when humidity is low (30–45% relative humidity), retaining much of the water that is necessary for high ionic conductivity (Fig. 1).

The nanocracks therefore have a similar role to stomatal pores in a cactus³. These pores open during cool, high-humidity periods at night and in the early morning, so that the plant can take up carbon dioxide with minimal water loss, but close when humidity is low (such as during the day) to limit water loss when it is hot. Similarly, the nanocracks on the membrane's surface close in low-humidity conditions, substantially decreasing water evaporation from the membrane, and so maintaining relatively high ionic conductivity.

The authors tested their surface-modified membranes in proton-exchange membrane fuel cells (PEMFCs) at elevated temperatures (greater than 100 °C) and low humidity. PEMFCs are a widely used class of fuel cell that generates electricity from hydrogen and

oxygen. The operation of PEMFCs at moderately high temperatures (100–150 °C) typically requires cumbersome equipment to control the temperature and humidity, but improves the tolerance of electrodes to impurities in the hydrogen-containing gas mixture that stop the fuel cell's catalyst from working at lower temperatures⁴.

Park *et al.* observed that their coated membranes displayed up to four times greater power density — a measure of the power (rate of energy conversion) that can be delivered per unit active area of the membrane — than uncoated analogues, with improvements most evident at 120 °C and low humidity. This performance enhancement partly stemmed from reduced water loss, but was also due to enhanced compatibility of the hydrophobic coating with the hydrophobic layers of catalyst in the PEMFC. The increased compatibility improved the cell's stability — in terms of both its performance and its physical stability — over periods of use of up to 220 hours at 120 °C, compared with a state-of-the-art, commercially available membrane.

Previous studies focused on modifying the bulk morphology of IEMs in the attempt to enhance low-humidity performance (see ref. 5, for example). Such approaches often involved introducing into the membrane hydrophobic components that cannot transport ions. In striking contrast to this, Park *et al.* attacked the problem at its source by modifying the membrane surface through which water loss occurs, thereby leaving the bulk properties of the membranes largely unchanged. Indeed, the

authors show that the surface-coating process does not alter either the membrane swelling at equilibrium conditions or the average spacing between the polymer chains through which ion transport occurs. Furthermore, the authors successfully applied their technique to different polymers and molecular architectures, suggesting that it might be applicable to a variety of membranes. Nevertheless, further study is needed to prove the generality of this surface-modification technique, its feasibility for scaling up and the long-term stability of the coated membranes.

The authors also explored the use of surface-modified membranes in reverse electrodialysis, a process that is used to harvest electricity from the energy released when two fluid streams of different salinity (such as river water and seawater) are mixed⁶. The membranes demonstrated a substantial increase in ion selectivity after surface modification, and ionic conductivity remained high — a combination of features that improves the energy efficiency of reverse electrodialysis. The authors' technique might therefore enhance membrane performance in other processes that require high selectivity and ion transport. Such surface modification has similarly been proposed⁷ to overcome the ubiquitous trade-off between throughput and selectivity that afflicts gas-separation membranes.

Park and colleagues report an excellent proof of concept for their technique, but many questions remain unanswered. A detailed understanding at the molecular level of the transport of water and ions in IEMs and of how polymer structure affects transport properties is only beginning to emerge⁸. Systematic studies of the authors' composite membranes will help to clarify this link between structure and transport, and will contribute to achieving the ultimate goal: the rational tailoring of many types of high-performance membrane for various applications. ■

Jovan Kamcev and Benny D. Freeman
are in the McKetta Department of Chemical Engineering, the Center for Energy and Environmental Resources, and the Texas Materials Institute, The University of Texas at Austin, Austin, Texas 78758, USA.
e-mails: jkamcev@utexas.edu;
freeman@che.utexas.edu

1. Park, C. H. *et al.* *Nature* **532**, 480–483 (2016).
2. Sata, T. *Ion Exchange Membranes: Preparation, Characterization, Modification and Application* (R. Soc. Chem., 2004).
3. Edwards, E. J. & Donoghue, M. J. *Am. Nat.* **167**, 777–793 (2006).
4. Li, Q., Jensen, J. O., Savinell, R. F. & Bjerrum, N. J. *Prog. Polym. Sci.* **34**, 449–477 (2009).
5. Chen, Y. *et al.* *Nature Chem.* **2**, 503–508 (2010).
6. Hong, J. G. *et al.* *J. Membr. Sci.* **486**, 71–88 (2015).
7. Robeson, L. M., Burgoyne, W. F., Langsam, M., Savoca, A. C. & Tien, C. F. *Polymer* **35**, 4970–4978 (1994).
8. Kamcev, J. *et al.* *Phys. Chem. Chem. Phys.* **18**, 6021–6031 (2016).

PALAEOONTOLOGY

Getting the measure of a monster

Scrutiny of fossils sometimes uncovers an unexpected phylogenetic relationship. New analyses of the enigmatic fossil *Tullimonstrum* from 300 million years ago reveal it to be a vertebrate. [SEE LETTERS P.496 & P.500](#)

SHIGERU KURATANI & TATSUYA HIRASAWA

What does it take to be called a monster? Having only ever been uncovered from one place, like the fish-like fossil *Palaeospondylus*¹ from a slate quarry in Scotland? Or displaying morphology that does not resemble any other animal, such as many of the fossils found in Canada's Burgess Shale? Nothing is more exciting in evolutionary zoology than to work out the taxonomic position of such 'monsters', because it has the potential to change the definition of animal phyla, or even an entire evolutionary scenario. The *Tullimonstrum* fossils of Illinois fall into this category: long known to palaeontologists², their identity has remained totally enigmatic. In two papers in this issue, McCoy *et al.*³ (page 496) and Clements *et al.*⁴ (page 500) now identify the 'Tully monster' as a vertebrate.

Different body plans define different animal phyla, and so the first step to identifying an animal is to identify its organ systems and their relative topographical relationships. However, for many fossils, not all organs are present, or their positions have been disturbed by post-mortem deformation. This is not the case for the Tully monster — a small animal, about 10 centimetres long², from the Carboniferous period (around 359 million to 299 million years ago). Despite thousands of fossils having been found in the one locality of Mazon Creek, Illinois, previous anatomical interpretations^{1,5} have been hampered by a lack of overt hard tissues (such as an exoskeleton) and the presence of some tremendously peculiar structures (Fig. 1).

A crucial advance made by McCoy and colleagues was the discovery of a notochord in *Tullimonstrum*. This flexible rod-like structure is the defining feature of all animals of the chordate group, which includes vertebrates. In *Tullimonstrum* fossils, the structure was previously interpreted as the gut⁵, but the animal has a true gut that looks similar to that in fossilized hagfish. On the basis of the reinterpretation of this structure as a notochord, McCoy *et al.* could also identify myomeres (segmented muscle blocks), a dorsal nerve cord, arcualia (paired cartilaginous nodules), a specialized head part, a tripartite brain-like structure and a tail fin consisting of dorsal and ventral

lobes. This set of basic features was found to be topographically arranged in a manner consistent with the vertebrate body plan.

But what about the Tully monster's main peculiarity — a stiff transverse bar with spherical structures at each end, previously interpreted² as eyes and stalk? This odd organ complex is unprecedented in vertebrates. Through a microscopic structural analysis and molecular analysis of compounds in the distal tip of the bar, Clements and colleagues conclude that the spherical organ could only be a vertebrate eye. In sum, the Tully monster possesses almost all the basic elements (in phylogenetic terms, the plesiomorphies) of a vertebrate.

If this is the case, which vertebrate taxon does this animal belong to? To answer that question, one has to identify certain derived features that are shared by only a specific group of vertebrates, such as the fur and three ear ossicles that are used to define the Mammalia. However, structures that are so derived and peculiar that they appear only in one group, such as the *Tullimonstrum* eye stalk, will not tell us about phylogenetic relationships. The Tully monster's proboscis, which is reminiscent of a similar structure in the *Opabinia*

fossils of around 525 million to 505 million years ago (during the Cambrian period), is another odd trait. In fact, it is these peculiarities that have kept this animal a monster for a long time.

On the tip of the proboscis is an oral apparatus that consisted of dorsal and ventral halves with horny teeth, which seem to have moved dorsoventrally. This structure looks similar to the jaw in tetrapods and most fishes. But the dorsoventral division does not necessarily indicate the presence of a jaw, because many agnathans (fossil and living jawless fishes), including the ammocoete larva of extant lampreys, have oral apparatuses with dorsal and ventral sections that can move against each other⁶.

According to McCoy *et al.*, the shaft of the proboscis was not an elastic tube but had an internal skeleton with joints³, resembling the lingual (tongue-like) apparatus of a cyclostome. The cyclostomes are the vertebrate lineage comprising the extant jawless fishes — lampreys and hagfishes. Other vertebrates are classed as gnathostomes (Fig. 1).

If the oral apparatus were really a jaw, it would be expected to resemble that of elasmobranchs (sharks, rays and skates) and sturgeons, whose upper jaw element (the palatoquadrate) is completely detached from the dorsal part of the skull. Yet, such an anatomical plan is inconsistent with the single nostril of the Tully monster, which has been presumed to be a mouth². Indeed, the presence of this single nostril excludes the monster from the extant gnathostome lineage. Furthermore, no paired fins have been identified in *Tullimonstrum*, which potentially places the animal even more basally on the animal tree than most of the gnathostomes.

McCoy *et al.* tentatively identify the animal as belonging to ancestral lampreys (Fig. 1). But

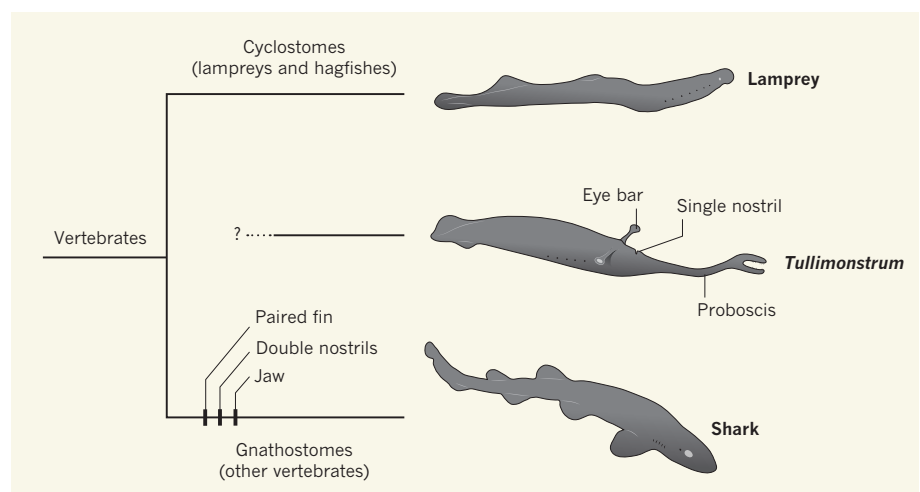


Figure 1 | Vertebrate descent. McCoy *et al.*³ and Clements *et al.*⁴ demonstrate that the *Tullimonstrum* fossils, which date to around 300 million years ago, were vertebrates. The vertebrates consist of two lineages, the cyclostomes and the gnathostomes, and McCoy *et al.* raise the possibility that *Tullimonstrum* evolved from the cyclostome lineage, in particular the lampreys. However, the unique features of the animal, including its proboscis, jaw-like oral apparatus and eye bars are still mysterious in vertebrate evolution.

because developmental studies have shown that lampreys and hagfishes are more closely related to each other than was previously imagined⁷, it could also represent an animal close to the origin of cyclostomes before the lamprey–hagfish divergence. That would mean that early cyclostomes exhibited a tremendous range of morphological variety. Indeed, the details of the early divergence of cyclostomes, which has been inferred to have occurred 390 million years

ago or before⁸, remain unclear. Perhaps even weirder fossil vertebrates remain to be dug up. ■

Shigeru Kuratani and Tatsuya Hirasawa are in the *Evolutionary Morphology Laboratory, RIKEN, Kobe, Hyogo 650-0047, Japan*.
e-mails: saizo@cdb.riken.jp;
hirasawa@cdb.riken.jp

1. Traquair, R. H. *Ann. Mag. Nat. Hist.* 6th ser. **6**, 479–486 (1890).

2. Johnson, R. G. & Richardson, E. S. *Fieldiana Geol.* **12**, 119–149 (1969).
3. McCoy, V. E. *et al. Nature* **532**, 496–499 (2016).
4. Clements, T. *et al. Nature* **532**, 500–503 (2016).
5. Richardson, E. S. *Science* **151**, 75–76 (1966).
6. Janvier, P. *Early Vertebrates* (Oxford Univ. Press, 1996).
7. Oisi, Y., Ota, K. G., Fujimoto, S. & Kuratani, S. *Nature* **493**, 175–180 (2013).
8. Kuraku, S. & Kuratani, S. *Zool. Sci.* **23**, 1053–1064 (2006).

This article was published online on 13 April 2016.

NUCLEAR PHYSICS

Four neutrons together momentarily

A system of four neutrons known as the tetra-neutron is a hypothetical state in nuclear physics. The report of evidence for the fleeting existence of this state has implications for research into neutron stars.

CARLOS A. BERTULANI
& VLADIMIR ZELEVINSKY

Atomic nuclei are composed of protons and neutrons, generically known as nucleons. These are not genuine elementary particles because they contain quarks and gluons, which interact with each other through the strong force (one of the four fundamental forces of nature). The strong interaction has subtle properties, with the most unsettling one being that quarks and gluons are never free, only confined within nucleons. Theorists continue to struggle to find exact solutions for various states of the highly complex quark–gluon systems, and to explain the nucleon–nucleon force that extends beyond

the confinement region. One long-sought state is the four-neutron system known as the tetra-neutron, which has no electric charge. Writing in *Physical Review Letters*, Kisamori *et al.*¹ present evidence for the existence of such a state.

In the authors' experiment, strongly bound α -particles (composed of two protons and two neutrons, and therefore identical to helium-4 nuclei) in liquid helium-4 (^4He) are used as a target for an incident beam of helium-8 (^8He , the 'projectile nucleus'). ^8He has two protons and six neutrons, and is produced in nuclear-fragmentation reactions in which oxygen-18 hits a beryllium target. The reaction between ^8He and ^4He is an appropriate choice for generating tetra-neutrons, because the four 'extra' neutrons in ^8He are weakly bound and

can be easily transformed in the interaction with ^4He .

The authors observed that the ^8He projectile exchanges two units of charge with the ^4He target and becomes a beryllium-8 nucleus (^8Be , four protons and four neutrons), the energy of which was measured with high precision. Because of charge conservation, the two protons in the target ^4He nucleus are substituted by neutrons, momentarily generating a four-neutron system in a quasi-bound state. This lasts only a few multiples of 10^{-22} seconds, after which it disassembles into free neutrons. This short-lived state appears as a bump in the energy spectrum of the ^8Be nucleus that emerges from the reaction.

Nuclear forces are essentially identical between all nucleons, whether they are protons or neutrons. So it might seem strange that the tetra-neutron is not bound but that the α -particle of two protons and two neutrons is strongly bound, despite the additional electrical repulsion between protons. The explanation is based on the Pauli exclusion principle, which forbids two identical nucleons from occupying the same quantum state. In the α -particle, all four particles can be in the same state because the two protons have opposite spins, as does the pair of neutrons, so that all four nucleons are different. But for four neutrons, only one pair can be in the lowest-energy state, forcing the second pair into a state of higher energy, thus making the tetra-neutron unstable.

By applying the principle of energy conservation to the studied nuclear reaction, Kisamori and colleagues infer that the tetra-neutron system has an internal excitation energy of about 0.8 million electronvolts (MeV); the excitation energy is the difference between the tetra-neutron mass and the mass of four free neutrons. If this quantity were less than zero, the system would be bound. For the observed tetra-neutron it is positive, making it an unbound system that exists for a short time before it decays into free neutrons. The statistical error (± 0.65 MeV) and systematic error (± 1.25 MeV) in the experiment are large, but the case for the existence of the tetra-neutron is compelling. The width of the bump in the ^8Be energy spectrum is about 2.6 MeV, and this energy uncertainty suggests that the state will eventually decay to another quantum state.

The hunt for the tetra-neutron has been going on for more than half a century, and

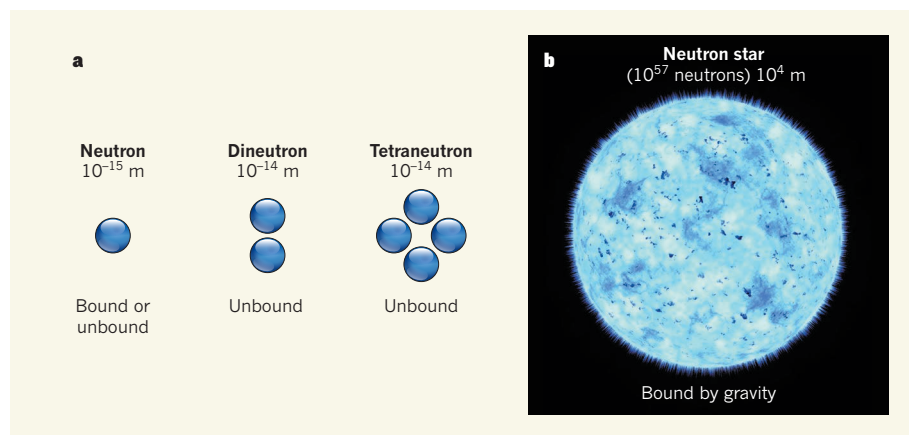


Figure 1 | Neutron systems. **a**, Neutrons have a radius of about one femtometre (10^{-15} m), and can be either bound in a nucleus or free (although unbound neutrons decay within about 15 minutes). Dineutrons, composed of two unbound neutrons, are ten times larger and unstable. Kisamori and co-workers' report evidence for a tetra-neutron (a system of four neutrons) that exists in a resonant state for about 10^{-22} seconds before dissociating into free neutrons. **b**, If the existence of the tetra-neutron state is confirmed, it will help to clarify nuclear interactions in few-nucleon systems, and possibly even in neutron stars.

experimentalists have announced the state's discovery before. In 2002, one collaboration claimed to have found a bound tetraneutron² in an experiment based on the detection of neutron clusters formed by fragmentation of beryllium-14 projectiles. But the result remains unconfirmed, and theorists quickly showed that, based on the best knowledge of the nucleon–nucleon interactions and other arguments^{3,4}, the existence of a bound tetraneutron was nearly impossible.

However, theorists could not rule out the existence of a tetraneutron as a short-lived 'resonant' state on the basis of a dineutron–dineutron structure^{3,4}. The dineutron state is formed by two neutrons, and is not stable. It is known as a virtual state: if its energy were reduced by 66 keV, then the dineutron system would become bound. Decades earlier, it had been proposed⁵ that dineutrons can become bound in the presence of additional nucleons; this mechanism is responsible for the properties of some bound nuclei that have a neutron excess, such as lithium-11, in which a pair of external neutrons forms a remote halo around the core of lithium-9.

The tetraneutron cannot form an atomic nucleus because it is charge neutral and therefore cannot hold electrons. But there is an intimate relationship between the tetraneutron structure and theoretical studies of neutron stars (Fig. 1), in which neutrons are compressed to densities more than 10^{14} times that of water⁶. They are prevented from imploding by an outward pressure that is

generated by the nucleon–nucleon interaction and other quantum-mechanical effects.

Nuclear physicists hope to develop a full understanding of how quarks and gluons inside nucleons generate nucleon–nucleon forces, and how many-body objects evolve to form complex structures such as the uranium nucleus and neutron stars. This is a formidable task, with well-understood parts but also many missing links. If Kisamori and co-workers' report of the tetraneutron state is confirmed, even as a short-lived resonance, it will add another structure to the nuclear chart that will help to improve our understanding of the nuclear interaction. ■

Carlos A. Bertulani is in the Department of Physics and Astronomy, Texas A&M University–Commerce, Commerce, Texas 75429-3011, USA. **Vladimir Zelevinsky** is in the Department of Physics and Astronomy, Michigan State University, East Lansing, Michigan 48824-1321, USA. e-mails: carlos.bertulani@tamuc.edu; zelevinsky@nscl.msu.edu

1. Kisamori, K. *et al.* *Phys. Rev. Lett.* **116**, 052501 (2016).
2. Marqués, F. M. *et al.* *Phys. Rev. C* **65**, 044006 (2002).
3. Bertulani, C. A. & Zelevinsky, V. J. *Phys. G* **29**, 2431–2437 (2003).
4. Pieper, S. C. *Phys. Rev. Lett.* **90**, 252501 (2003).
5. Migdal, A. B. *Sov. J. Nucl. Phys.* **16**, 238–241 (1973).
6. Shapiro, S. L. & Teukolsky, S. A. *Black Holes, White Dwarfs, and Neutron Stars* (Wiley, 1983).

This article was published online on 6 April 2016.

NEUROSCIENCE

Fault tolerance in the brain

If stored information is erased from neural circuits in one brain hemisphere in mice, the lost data can be recovered from the other. This finding highlights a safeguarding mechanism at work in the brain. [SEE ARTICLE P.459](#)

BYRON M. YU

When we send an e-mail or save a file on our hard drives, information can be lost, owing to dropped data packets or corrupted bits. We typically do not notice such failures because systems are designed with built-in mechanisms to restore the lost data. Dropped packets are retransmitted, and multiple copies of data are saved. The brain also stores and transmits information — is it, too, fault-tolerant? In this issue, Li *et al.*¹ (page 459) report the perturbation of brain activity to erase stored information in mice. They discover that the lost information can

be rapidly restored by an unperturbed brain region.

The brain can reorganize itself to restore function after certain types of injury², but this type of fault tolerance typically takes place over weeks. By contrast, many everyday brain functions, such as putting a name to the face of an acquaintance or hitting a tennis ball, take place on a timescale of seconds or less. Does a fault-tolerance mechanism also operate in neural circuits over these shorter timescales?

Li *et al.* investigated whether regions present in each of the brain's two hemispheres might act together to produce a rapid back-up system for stored information — a mechanism



50 Years Ago

Hypnotic Susceptibility. By Ernest R. Hilgard — A large number of studies designed to investigate various hypnotic phenomena have been carried out by Ernest Hilgard and his co-workers on a considerable number of college students during the past eight years. Individual differences in 'hypnotizability' have been a major area of interest and in the course of their investigations several scales were developed for the quantitative assessment of hypnotic susceptibility ... There are three general purpose scales and a scale for yielding profiles of hypnotic ability. Convincing statistical evidence is given concerning their validity and reliability ... The latter part of the book is concerned with the relation of hypnotic susceptibility to a number of personality variables ... Although some significant correlations do emerge, they are insufficient to characterize the hypnotizable person clearly.

From *Nature* 30 April 1966

100 Years Ago

The large meteors which passed over Northern America on February 9, 1913, presented some unique features. The length of their observed flight was about 2600 miles, and they must have been moving in paths concentric, or nearly concentric, with the earth's surface, so that they temporarily formed new terrestrial satellites ... The meteors were last seen from the Bermuda Islands ... I have since made efforts to obtain further observations from seafaring men through the medium of the *Nautical Magazine*, and have succeeded in procuring data which prove that the meteors were observed during a course of 5500 miles from about lat. 51° N., long. 107° W., to lat. 5½° S., long. 32½° W.

From *Nature* 27 April 1916

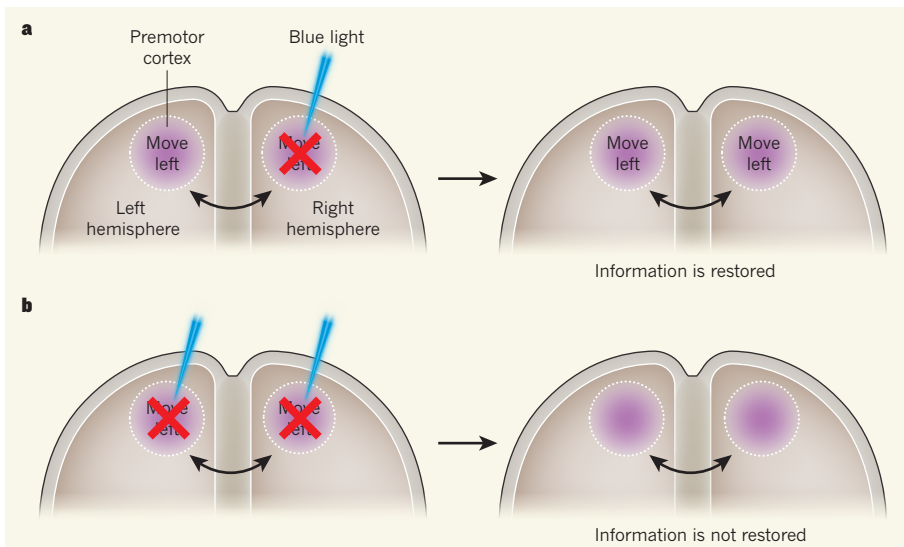


Figure 1 | A mechanism of redundancy. **a**, The premotor cortex regions in each hemisphere of the mouse brain, which are connected by neurons (double-headed arrows), produce activity that prepares the animal to move its tongue left or right. Li *et al.*¹ genetically engineered mice such that blue light could temporarily block neuronal activity in these regions, erasing information about the intended direction of movement. If information is erased in one hemisphere, it is quickly restored. **b**, By contrast, if information is erased in both hemispheres, it is not restored.

known as redundancy. Specifically, they tested whether the two premotor cortices of the mouse brain act redundantly to prepare the animal to lick with its tongue in a particular direction, which it has been taught will lead to a reward of water. The authors briefly blocked the activity of premotor neurons in one hemisphere and observed that information about intended licking direction was quickly restored (Fig. 1a). However, if they silenced neurons in both hemispheres in this way, the information was not restored, and so the animal licked left or right at random (Fig. 1b).

These results indicate that, during single-hemisphere silencing, fault tolerance is provided by the unmodified hemisphere. To test this back-up system more directly, the researchers severed the connections between the two hemispheres. When neurons in one hemisphere were silenced in this setting, the information about intended licking direction was not restored.

Next, Li *et al.* constructed computational network models of neurons in two interacting hemispheres to study how connectivity between the two brain regions enables fault tolerance. In these models, as in the experimental setting, information about movement direction was restored after neuronal silencing in one hemisphere. Together with the experimental evidence, these data suggest that each hemisphere helps the other to restore information about planned movement direction.

Perhaps the most interesting finding in this study is that, after silencing neurons in one hemisphere, not all aspects (called dimensions) of the perturbed neural activity recovered equally. Li and colleagues found that the neural activity that enabled maximal differentiation

between left and right licks recovered rapidly. By contrast, other dimensions of neural activity that were not relevant to the task did not always recover. Thus, there was preferential recovery of the dimension that was needed for the animal to succeed at the licking task.

The current study involved both hemispheres controlling a single effector, the tongue. An open question is how these findings apply to brain functions that predominantly involve a single hemisphere, such as control over reaching with one arm. As the authors point out, one possibility is that there are redundant subcircuits within a hemisphere, perhaps spread across multiple brain areas, working together to provide fault tolerance.

Li *et al.* perturbed neural activity using an

optogenetic technique, in which the activity of neurons that harbour light-sensitive ion channels can be modulated using light. This approach allows the silencing or activation of many neurons in unison. To further understand the fault-tolerant properties of neural circuits, more-flexible methods that allow selective activation and silencing of different groups of neurons at different times are needed. Such methods would permit testing of the robustness of a neural circuit to different patterns of perturbation, including those that mimic the random signal disturbances, known as noise, that are a part of normal neuronal signalling³.

The current work demonstrates the power of perturbing neural activity in combination with multidimensional analysis of the activity of a neural population⁴. By perturbing neural activity in different ways and observing how it recovers, we should be able to gain further insights into fundamental network-level mechanisms that support brain functions⁵. Advances in methods for perturbing and recording neural activity, for analysing population-wide neural activity and for network modelling are rapidly making such studies possible. ■

Byron M. Yu is in the Department of Electrical and Computer Engineering and the Department of Biomedical Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA. e-mail: byronyu@cmu.edu

1. Li, N., Daie, K., Svoboda, K. & Druckmann, S. *Nature* **532**, 459–464 (2016).
2. Kolb, B. & Whishaw, I. Q. *Prog. Neurobiol.* **32**, 235–276 (1989).
3. Faisal, A. A., Selen, L. P. J. & Wolpert, D. M. *Nature Rev. Neurosci.* **9**, 292–303 (2008).
4. Cunningham, J. P. & Yu, B. M. *Nature Neurosci.* **17**, 1500–1509 (2014).
5. Brody, C. D., Romo, R. & Kepecs, A. *Curr. Opin. Neurobiol.* **13**, 204–211 (2003).

This article was published online on 13 April 2016.

MATHEMATICAL PHYSICS

Glitches in time

A mathematical technique has now been developed that reveals the underlying dynamics of time-dependent data collected with extreme temporal uncertainty, without using additional, costly instrumentation. [SEE LETTER P.471](#)

CHARLOTTE A. L. HALEY

Many of today's scientific questions require the reconstruction of accurate histories from data collected with uncertainty in the temporal or spatial measurement process. Examples include the systematic errors in timing or spacing seen in measurements of astronomical

time series collected at uneven intervals^{1,2}; in determinations of global climate history from concentrations of gas extracted from bubbles in ice cores drilled in the Antarctic³; and in deductions of the proximity of far-away galaxies when the observed light is bent by gravitational lensing⁴. On page 471 of this issue, Fung *et al.*⁵ report a mathematical method that allowed them to analyse data with uneven temporal

spacing or samples that appeared out of order.

Free-electron lasers allow the early stages of ultrafast chemical reactions to be studied with molecular precision^{6,7}. Fung *et al.* used X-ray free-electron lasers to investigate a fundamental chemical process: the photoionization of nitrogen molecules, whereby the nitrogen ions N_2^+ and N_2^{2+} are created after photon absorption by molecular nitrogen (N_2). The ionization plays out incredibly quickly, over a period of several hundred femtoseconds (1 fs is 10^{-15} seconds).

The authors used an infrared pulse to trigger photoionization and an X-ray pulse to probe the process, capturing snapshots using a spectrometer with a 'shutter speed' of less than 100 fs. Unfortunately, thermal noise in the instrument introduces timing uncertainty that can exceed 280 fs (ref. 7). This temporal uncertainty, or jitter, makes it difficult to measure quantities such as the vibrating frequencies of the molecular system.

Fung *et al.* used a mathematical technique known as nonlinear Laplacian spectral analysis (NLSA) to recover the true dynamics underlying these snapshots (Fig. 1). In this technique, the data are embedded in a multi-dimensional space characterized by Laplacian functions, which can represent turbulent or intermittent structure in the data. The technique is nonlinear because the data are mapped to a curved surface. The term 'spectral' is used because, just as a prism separates white light into a spectrum of long to short wavelengths, NLSA decomposes the snapshots into their constituent time and space components (known as chronograms and topograms, respectively), arranged from strongest (the largest contributors to the signal) to weakest (the smallest contributors; Fig. 1b,c). The authors' key observation is that the chronograms are robust to jitter. In other words, the data points in the chronograms are uniformly spaced in time, which means that it is, in principle, possible to reconstruct the original data without jitter by recombining the modes.

However, NLSA does not allow such a reconstruction except in simple examples, because it requires a huge number of unknown parameters to be estimated. But reconstruction is not the goal. Instead, the quantities of interest can be observed directly by looking at the strongest 'modes' obtained from NLSA — these modes are derived mathematically from the chronograms and topograms, and describe the essential properties of the nitrogen system. One of the modes reveals an especially interesting result: molecular nitrogen vibrates with three previously unobserved frequencies. These are in the 16–24-fs range — much shorter than those observed in any preceding experiment, but consistent with predictions from quantum mechanical theory.

Aside from this important application to quantum molecular physics, the theory proposed by Fung *et al.* has profound implications

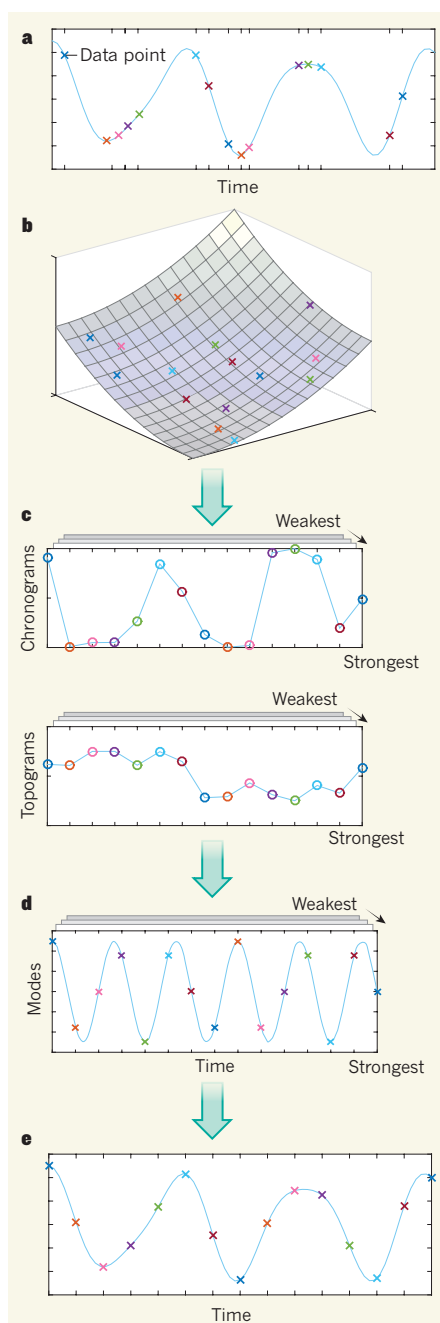


Figure 1 | A mathematical process for analysing space-time data. Fung *et al.*⁵ studied the vibrations and break-up of nitrogen molecules in response to a single infrared trigger pulse and a probing X-ray pulse. **a**, The collected space-time data were unevenly distributed in time and sometimes in the wrong order (data from a single spatial coordinate are shown). **b**, **c**, To extract the underlying temporal dynamics, the authors embedded the data in a five-dimensional nonlinear space (**b**, three dimensions are shown, for simplicity), then decomposed the data into a series of time and space components (**c**, chronograms and topograms, respectively), organized from strongest to weakest. **d**, These components can be used to construct modes that describe the essential properties of the nitrogen system and have uniform time spacing. **e**, In principle, the original data without timing errors can be reconstructed from a combination of these modes.

for signal processing and spatiotemporal statistics. Techniques for the spectral analysis of time series or spatial data involving uneven sampling or missing data, both with known and unknown timestamps, have been sought since at least the early 1980s (refs. 2,3,8,9). It is remarkable that, in the molecular-nitrogen experiment, the authors can resolve the vibrational modes to about 1 fs spacing. However, an expression for the reduction in temporal uncertainty achieved after application of this technique is desirable. Another question is whether NLSA can return a uniform result if signals are systematically bunched in time. And finally, if NLSA is applicable to unevenly timed samples, it ought to be applicable to unevenly spaced samples in one spatial dimension — but are there nonlinear embeddings that can extend this result into higher dimensions?

The authors verified their observations by numerical simulation, but mathematical rigour and theoretical formalism for the temporal uniformity of the chronogram components are still lacking. Moreover, the presented formulation may still be overly complex. In particular, the authors show that a single cosine curve, sampled with jitter, can be recovered without the need for a nonlinear approach. Given that this simple cosine example underlies all conventional spectrum analysis¹⁰, is there a simpler formulation that does not require the complex machinery of NLSA but still possesses the property of time uniformity in the result? What is the simplest, most basic working example of such a process? This work opens the door to many interesting questions. ■

Charlotte A. L. Haley is in the Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, Illinois 60439, USA.
e-mail: haley@anl.gov

1. Scargle, J. D. *Astrophys. J.* **263**, 835–853 (1982).
2. Kuhn, J. R. *Astron. J.* **87**, 196–202 (1982).
3. Petit, J. R. *et al. Nature* **399**, 429–436 (1999).
4. Wittman, D. M., Tyson, J. A., Kirkman, D., Dell'Antonio, I. & Bernstein, G. *Nature* **405**, 143–148 (2000).
5. Fung, R. *et al. Nature* **532**, 471–475 (2016).
6. Pellegrini, C. & Stöhr, J. *Nucl. Instrum. Methods Phys. Res. A* **500**, 33–40 (2003).
7. Glowina, J. M. *et al. Opt. Express* **18**, 17620–17630 (2010).
8. Kondrashov, D. & Ghil, M. *Nonlin. Process. Geophys.* **13**, 151–159 (2006).
9. Fuentes, M. *J. Am. Stat. Assoc.* **102**, 321–331 (2007).
10. Ghil, M. *et al. Rev. Geophys.* **40**, 3–1–3–41 (2002).

CORRECTION

In the News & Views article 'Physics: Quantum problems solved through games' by Sabrina Maniscalco (*Nature* **532**, 184–185; 2016), reference 7 was incorrect. The correct reference is <https://www.scienceathome.org/games/quantum-moves/game>

Natural speech reveals the semantic maps that tile human cerebral cortex

Alexander G. Huth¹, Wendy A. de Heer², Thomas L. Griffiths^{1,2}, Frédéric E. Theunissen^{1,2} & Jack L. Gallant^{1,2}

The meaning of language is represented in regions of the cerebral cortex collectively known as the ‘semantic system’. However, little of the semantic system has been mapped comprehensively, and the semantic selectivity of most regions is unknown. Here we systematically map semantic selectivity across the cortex using voxel-wise modelling of functional MRI (fMRI) data collected while subjects listened to hours of narrative stories. We show that the semantic system is organized into intricate patterns that seem to be consistent across individuals. We then use a novel generative model to create a detailed semantic atlas. Our results suggest that most areas within the semantic system represent information about specific semantic domains, or groups of related concepts, and our atlas shows which domains are represented in each area. This study demonstrates that data-driven methods—commonplace in studies of human neuroanatomy and functional connectivity—provide a powerful and efficient means for mapping functional representations in the brain.

Previous neuroimaging studies have identified a group of regions that seem to represent information about the meaning of language. These regions, collectively known as the semantic system, respond more to words than non-words¹, more to semantic tasks than phonological tasks¹, and more to natural speech than temporally scrambled speech². Studies that have investigated specific types of representation in the semantic system have found areas selective for concrete or abstract words^{3–5}, action verbs⁶, social narratives⁷ or other semantic features. Others have found areas selective for specific semantic domains—groups of related concepts such as living things, tools, food or shelter^{8–13}. However, all previous studies tested only a handful of stimulus conditions, so no study has yet produced a comprehensive survey of how semantic information is represented across the entire semantic system.

We addressed this problem by using a data-driven approach¹⁴ to model brain responses elicited by naturally spoken narrative stories that contain many different semantic domains¹⁵. Seven subjects listened to more than two hours of stories from *The Moth Radio Hour*² while whole-brain blood-oxygen-level-dependent (BOLD) responses were recorded by fMRI. We then used voxel-wise modelling, a highly effective approach for modelling responses to complex natural stimuli^{14–17}, to estimate the semantic selectivity of each voxel (Fig. 1a).

Voxel-wise model estimation and validation

In voxel-wise modelling, features of interest are first extracted from the stimuli and then regression is used to determine how each feature modulates BOLD responses in each voxel. We used a word embedding space to identify semantic features of each word in the stories^{12,15,18–20}. The embedding space was constructed by computing the normalized co-occurrence between each word and a set of 985 common English words (such as ‘above’, ‘worry’ and ‘mother’) across a large corpus of English text. Words related to the same semantic domain tend to occur in similar contexts, and so have similar co-occurrence values. For example, the words ‘month’ and ‘week’ are very similar (the correlation between the two is 0.74), while the words ‘month’ and ‘tall’ are not (correlation –0.22).

Next we used regularized linear regression to estimate how the 985 semantic features influenced BOLD responses in every cortical voxel and in each individual subject (Fig. 1a). To account for responses

caused by low-level properties of the stimulus such as word rate and phonemic content, additional regressors were included during voxel-wise model estimation and then discarded before further analysis. We also included additional regressors to account for physiological and emotional factors, but these had no effect on the estimated semantic models (Supplementary Data 3).

One advantage of voxel-wise modelling over conventional neuroimaging approaches is that the fit models can be validated by predicting BOLD responses to new natural stimuli that were not used during model estimation. This makes it possible to compute effect size by finding the fraction of response variance explained by the models. We tested how well the voxel-wise models predicted BOLD responses elicited by a new 10-min *Moth* story (Fig. 1b) that had not been used for model estimation. We found good prediction performance for voxels located throughout the semantic system, including in the lateral temporal cortex (LTC) and ventral temporal cortex (VTC), lateral parietal cortex (LPC) and medial parietal cortex (MPC), and medial prefrontal cortex, superior prefrontal cortex (SPFC) and inferior prefrontal cortex (IPFC) (Fig. 1c and Extended Data Fig. 1). This suggests that much of the semantic system is domain selective.

Mapping semantic representation across cortex

By inspecting the fit models, we can determine which specific semantic domains are represented in each voxel. In theory this could be done by examining each voxel separately. However, our data consist of tens of thousands of voxels per subject, rendering this approach unfeasible. A practical alternative is to project the models into a low-dimensional subspace that retains as much information as possible about the semantic tuning of the voxels^{10,14}. We found such a space by applying principal components analysis to the estimated models aggregated across subjects, producing 985 orthogonal semantic dimensions that are ordered by how much variance each explained across the voxels. It is likely that only some of these dimensions capture shared aspects of semantic tuning across the subjects; the rest reflect individual differences, fMRI noise, or the statistical properties of the stories. To identify the shared dimensions, we tested whether each explained more variance across the models than expected by chance, which was defined by the principal components of the stimulus matrix used for model estimation¹⁴. At least four dimensions explained a significant amount

¹Helen Wills Neuroscience Institute, University of California, Berkeley, California 94720, USA. ²Department of Psychology, University of California, Berkeley, California 94720, USA.

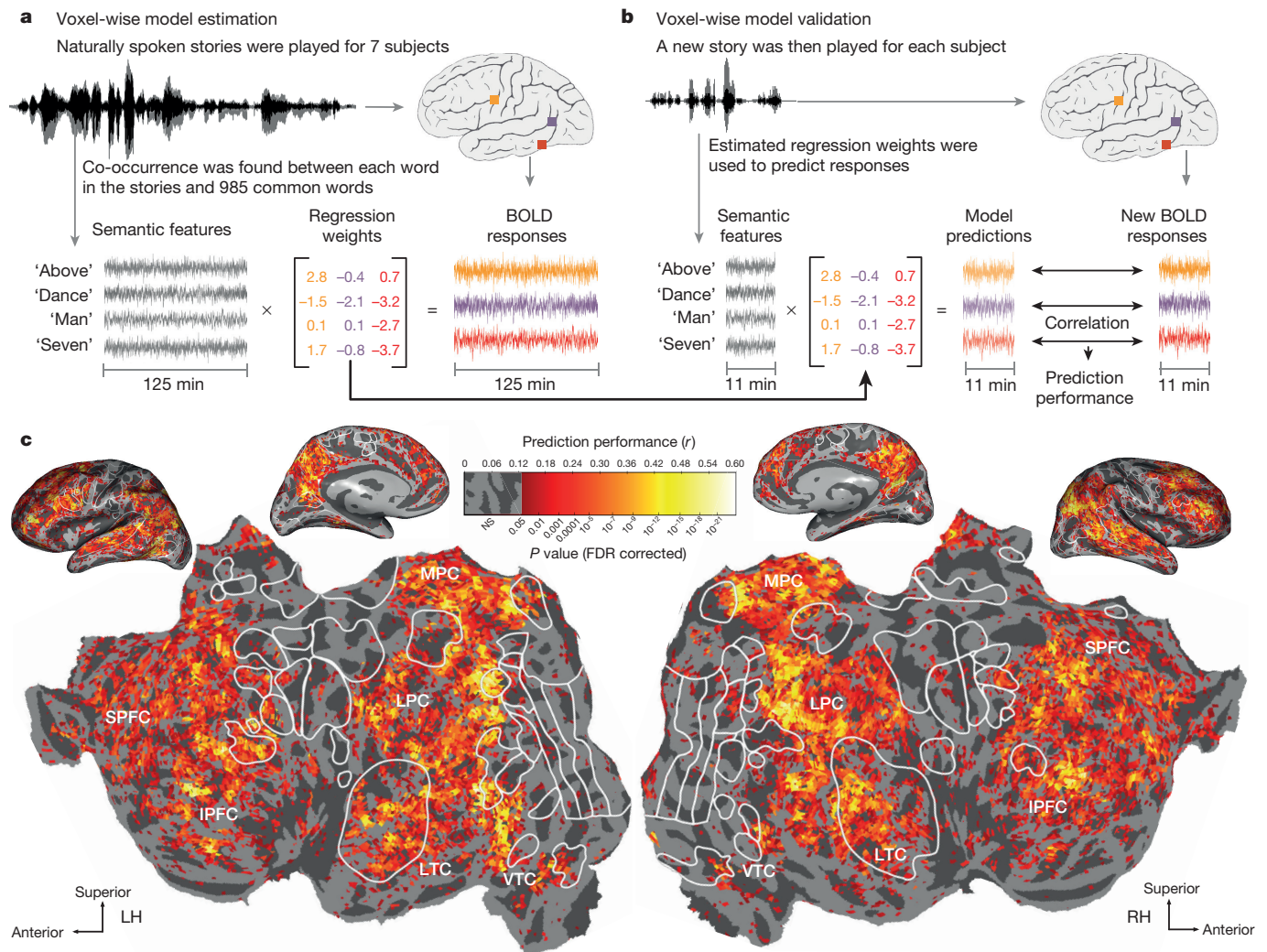


Figure 1 | Voxel-wise modelling. **a**, Seven subjects listened to over 2 h of naturally spoken narrative stories while BOLD responses were measured using fMRI. Each word in the stories was projected into a 985-dimensional word embedding space constructed using word co-occurrence statistics from a large corpus of text. A finite impulse response (FIR) regression model was estimated individually for every voxel. The voxel-wise model weights describe how words appearing in the stories influence BOLD signals. **b**, Models were tested using one 10-min story that was not

of variance ($P < 0.001$, Bonferroni-corrected bootstrap test) in all but one subject; in the last subject only three dimensions were significant (Extended Data Fig. 2). This suggests that our fMRI data contain about four statistically significant semantic dimensions that are shared across subjects.

The four shared semantic dimensions provide a way to summarize succinctly the semantic selectivity of a voxel. However, to interpret projections of the models onto these dimensions we need to understand how semantic information is encoded in this four-dimensional space. To visualize the semantic space, we projected the 10,470 words in the stories from the word embedding space onto each dimension. We then used *k*-means clustering to identify 12 distinct categories (see Supplementary Methods for details). Each category was inspected and labelled by hand. The labels assigned to the 12 categories were 'tactile' (a cluster containing words such as 'fingers'), 'visual' (words such as 'yellow'), 'numeric' ('four'), 'locational' ('stadium'), 'abstract' ('natural'), 'temporal' ('minute'), 'professional' ('meetings'), 'violent' ('lethal'), 'communal' ('schools'), 'mental' ('asleep'), 'emotional' ('despised') and 'social' ('child'). (See Supplementary Table 2 and Supplementary Data 5 for more detailed evaluations of each category.)

included during model estimation. Model prediction performance was computed as the correlation between predicted responses to this story and actual BOLD responses. **c**, Prediction performance of voxel-wise models for one subject. Semantic models accurately predict BOLD responses in many brain areas, including the LTC, VTC, LPC, MPC, SPFC and IPFC. These regions have previously been identified as the semantic system in the human brain. LH, left hemisphere; RH, right hemisphere.

Next, we visualized where each of the 12 categories appeared in the shared semantic space (Fig. 2a). Each category label was also assigned an RGB colour, where the red channel was determined by the first dimension, the green channel by the second, and the blue channel by the third. The first dimension is that which captured the most semantic variance across the voxel-wise models of all seven subjects. One end of this dimension favours categories related to humans and social interaction, including 'social', 'emotional', 'violent' and 'communal'. The other end favours categories related to perceptual descriptions, quantitative descriptions and setting, including 'tactile', 'locational', 'numeric' and 'visual'. This is consistent with previous suggestions that humans comprise a particularly salient and strongly represented semantic domain^{16,21}. Subsequent dimensions of the semantic space captured less variance than the first and were also more difficult to interpret. The second dimension seems to distinguish between perceptual categories, including 'visual' and 'tactile', and non-perceptual categories, including 'mental', 'professional' and 'temporal'. The third and fourth dimensions are less clear.

Earlier studies identified the cortical regions comprising the semantic system^{1,2}, but could not comprehensively characterize their semantic

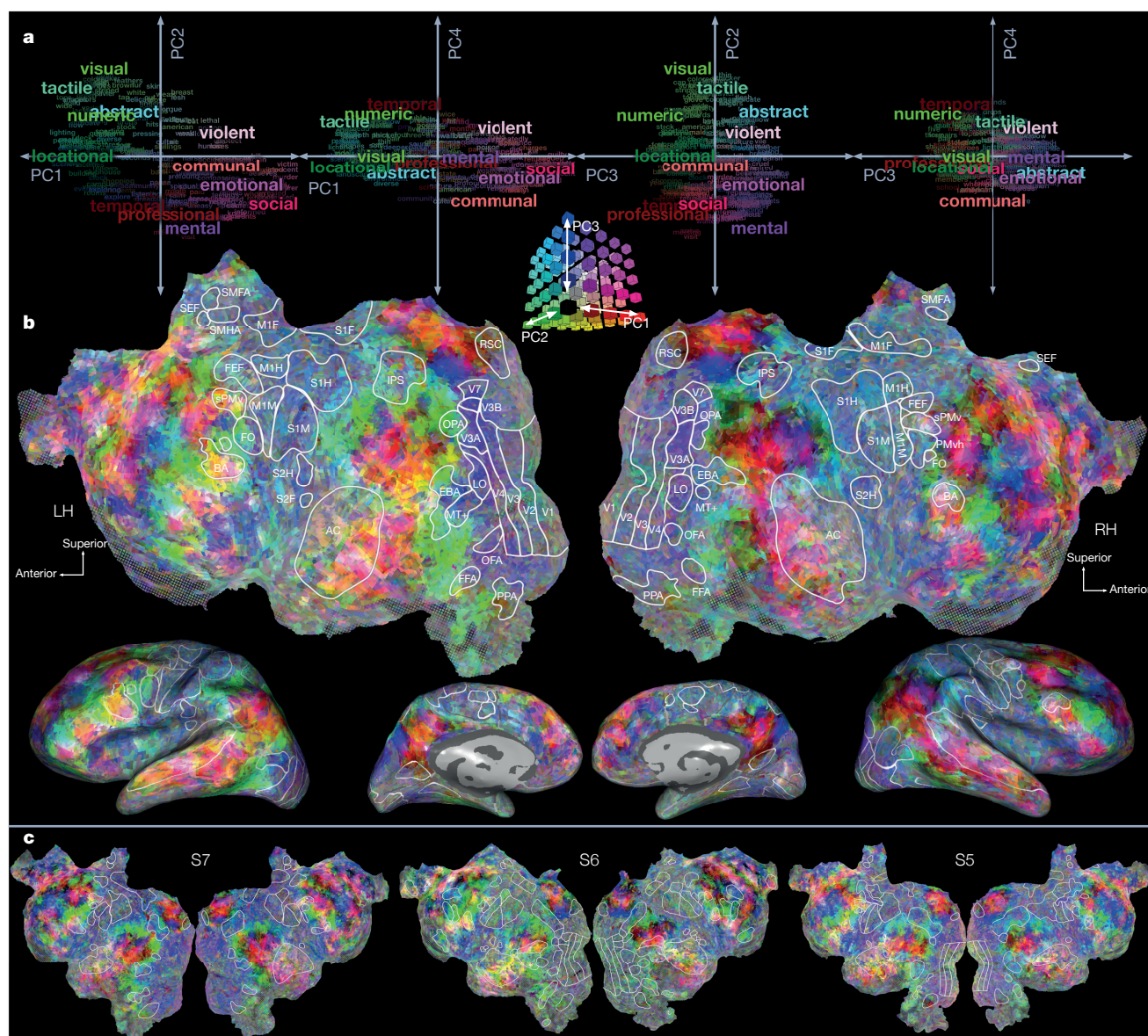


Figure 2 | Principal components of voxel-wise semantic models. **a–c**, Principal components analysis of voxel-wise model weights reveals four important semantic dimensions in the brain (Extended Data Fig. 2). **a**, An RGB colourmap was used to colour both words and voxels based on the first three dimensions of the semantic space. Words that best match the four semantic dimensions were found and then collapsed into 12 categories using *k*-means clustering. Each category (Supplementary Table 2) was manually assigned a label. The 12 category labels (large words) and a selection of the 458 best words (small words) are plotted here along four pairs of semantic dimensions. The largest axis of variation lies roughly along the first dimension, and separates perceptual and physical

categories (tactile, locational) from human-related categories (social, emotional, violent). PC, principal component. **b**, Voxel-wise model weights were projected onto the semantic dimensions and then coloured using the same RGB colourmap (see Extended Data Fig. 3 for separate dimensions). Projections for one subject (S2) are shown on that subject's cortical surface. Semantic information seems to be represented in intricate patterns across much of the semantic system. **c**, Semantic principal component flatmaps for three other subjects. Comparing these flatmaps, many patterns appear to be shared across individuals. (See Extended Data Fig. 3 for other subjects.) Abbreviations for regions of interest are listed in the Methods section.

selectivity. We were able to visualize the pattern of semantic-domain selectivity across the entire cortex by projecting voxel-wise models onto the shared semantic dimensions. Figure 2b shows projections onto the first three dimensions for one subject, plotted together using the same RGB colour scheme as in Fig. 2a (Extended Data Fig. 3a shows each dimension separately). Thus, for example, a green voxel produces greater BOLD responses to categories that are coloured green in the semantic space, such as 'visual' and 'numeric'. This visualization suggests that semantic information is represented in intricate patterns that cover the semantic system, including broad regions of the prefrontal cortex, LTC and MTC, and LPC and MPC. Furthermore, these patterns appear to be relatively consistent across individuals (Fig. 2c; see also Extended Data Fig. 3b).

Using PrAGMATiC to construct a semantic atlas

Given the apparent consistency in the patterns of semantic selectivity across individuals, we sought to create a single atlas that describes the distribution of semantically selective functional areas in human cerebral cortex. To accomplish this, we developed a new Bayesian algorithm, PrAGMATiC, that produces a probabilistic and generative model of areas tiling the cortex²². This algorithm models patterns of functional tuning recovered by voxel-wise modelling as a dense, tiled map of functionally homogeneous brain areas (Fig. 3a), while respecting individual differences in anatomical and functional anatomy^{23,24}. The arrangement and selectivity of these areas are determined by parameters learned from the fMRI data through a maximum-likelihood estimation technique similar to contrastive divergence²⁵.

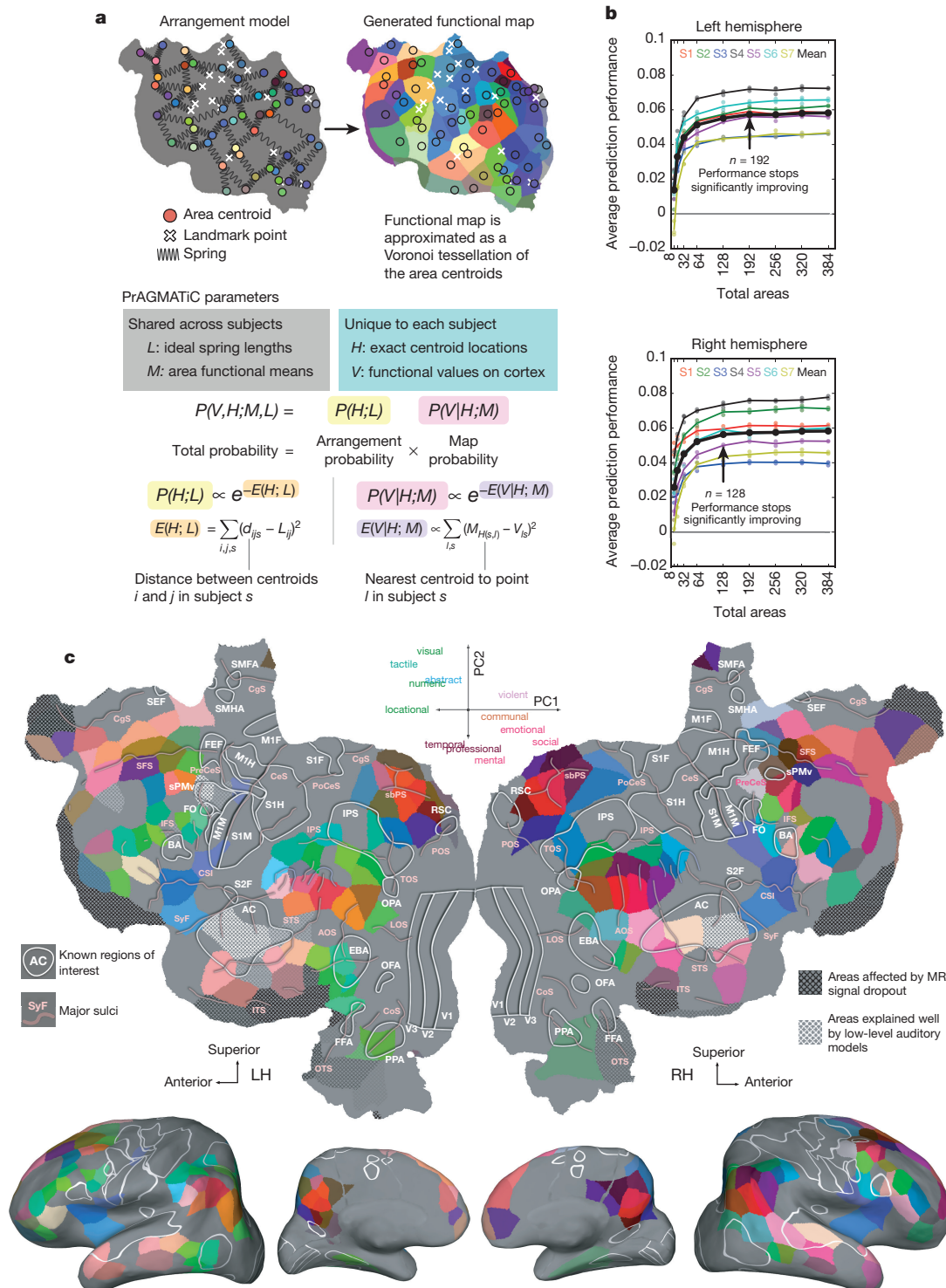


Figure 3 | PrAGMATiC: a generative model for cortical maps.

a–c. To create an atlas that describes the distribution of semantically selective functional areas in the human cerebral cortex we developed PrAGMATiC, a probabilistic and generative model of areas tiling the cortex. **a.** PrAGMATiC has two parts: an arrangement model and an emission model. The arrangement model is analogous to a physical system of springs joining neighbouring area centroids. To enforce similarity across subjects, springs also join areas to 19 regions of interest that were localized separately. The emission model assigns the functional mean of the closest area centroid to each point on the cortex, forming a Voronoi tessellation. Spring lengths and area means are shared across subjects while exact area locations are unique to each subject. These parameters are fit using maximum-likelihood estimation. **b.** A leave-one-out procedure

was used to choose the number of areas in each hemisphere. PrAGMATiC models were estimated on six subjects and then used to predict BOLD responses for the seventh. Prediction performance improved significantly up to 192 total areas in the left hemisphere and 128 areas in the right. **c.** A semantic atlas was estimated using data from all seven subjects. Areas for which the semantic model did not predict better than models based on low-level features (that is, word rate, phonemes) were removed. The remaining areas were plotted on one subject's cortical surface using the same RGB colourmap as Fig. 2. Areas dominated by signal dropout are shown in black hatching, and areas where the low-level models performed well are shown in white hatching. This atlas shows the functional organization of the semantic system that is common across subjects.

Some parameters are shared; these describe properties of the cortical map that are common across the group. Other parameters are unique to each subject; these capture individual differences. Learning both shared and unique parameters simultaneously eliminates the usual requirement to perform anatomical or functional alignment of data across subjects.

The PrAGMATiC algorithm has two components: an arrangement model that determines where functional areas appear on the cortical sheet, and an emission model that determines how the cortical map is produced from an arrangement of areas. The arrangement model simulates a physical spring network that joins the centroid of each functional area to its neighbours. Equilibrium spring lengths are shared across subjects, but each spring can be stretched or compressed in any individual subject. Arrangements are also constrained by several functional landmarks, which are known regions of interest identified in every subject using separate functional data. These constraints ensure that the maps will be similar across subjects, but allow for substantial individual variability in the precise arrangement and size of the areas. Using the arrangement model, the emission model creates homogeneous functional areas by assigning each vertex on the cortical surface to the nearest area centroid. The functional value at each vertex is then drawn from a multivariate normal distribution. The mean functional value for each area is learned by the algorithm and is shared across subjects. We define the functional value as a four-dimensional vector that reflects the projection of the estimated model for each voxel onto the four shared semantic dimensions.

One important hyperparameter is the total number of areas that PrAGMATiC uses to tile the cortex. We used a cross-validation procedure to choose the total number of areas tiling each hemisphere and then tested whether each area is semantically selective. PrAGMATiC models were estimated from data from six subjects and then used to predict the semantic map in the seventh subject using only cortical anatomy and the locations of functional landmarks in that subject. Predicted BOLD responses based on this map were compared to actual responses to determine how well the PrAGMATiC model generalizes across subjects. Prediction performance climbed quickly as the total number of areas rose from 8 to 128 and improved more gradually thereafter (Fig. 3b). In the left hemisphere, prediction performance did not improve significantly for models with 192 or more total areas (false discovery rate (FDR) > 0.01, Tukey post-hoc test with subject-wise random effects). In the right hemisphere, prediction performance did not improve significantly for models with 128 or more total areas. However, because PrAGMATiC tiles the entire cerebral cortex, these numbers include both semantically selective and nonselective areas. To identify the semantically selective areas and eliminate those that are nonselective, we tested whether the average voxel-wise semantic model in each area predicted responses significantly better than the average model for low-level features such as word rate, phoneme rate, and phonemes. This excluded areas that were not selective for either semantic or low-level features, such as motor and visual cortex. It also excluded areas that were not uniquely selective for semantic features, such as Broca's area, which was desirable because of the increased uncertainty of semantic model weights in those areas.

Figure 3c shows the semantic atlas projected onto the cortical surface of one subject (see also Extended Data Figs 4 and 5). The left hemisphere contains 77 semantic areas (FDR < 1/192, bootstrap test) and the right contains 63 semantic areas (FDR < 1/128, bootstrap test). A diverse tiling of areas that represent different semantic domains appear in the LPC (Extended Data Fig. 6), MPC (Extended Data Fig. 7) and SPFC (Extended Data Fig. 8). In the LPC and MPC, central areas (near the angular gyrus and subparietal sulci, respectively) are selective for social concepts, while surrounding areas are selective for numeric, visual or tactile concepts. In the SPFC, medial areas are mainly selective for social concepts, while dorsolateral areas are more diverse. The LPC, MPC and SPFC also all belong to the default mode network (DMN), which is thought to be involved in introspection, rumination

and conscious thought²⁶. One interesting possibility is that the semantic areas identified here represent the same semantic domains during conscious thought. This suggests that the contents of thought, or internal speech, might be decoded using these voxel-wise models¹⁷. In the LTC (Extended Data Fig. 9), our atlas identifies fewer distinct semantic areas than in the LPC, MPC or SPFC. This is surprising because the LTC has a key role in language comprehension^{1,27} and also belongs to the DMN. However, the quality of fMRI signals recorded in the anterior temporal lobe is poor, so the LTC probably contains other semantic areas that could not be recovered using our current approach. Detailed analyses of semantic representations in the LPC, MPC, SPFC and LTC, as well as the VTC (Extended Data Fig. 10), IPFC (Extended Data Fig. 11), and opercular and insular cortex (Extended Data Fig. 12) can be found in Supplementary Information, along with discussion and comparisons to earlier neuroimaging and lesion results.

Discussion

One striking aspect of our atlas is that the distribution of semantically selective areas is relatively symmetrical across the two cerebral hemispheres. This finding is inconsistent with human lesion studies that support the idea that semantic representation is lateralized to the left hemisphere¹³. However, many fMRI studies of semantic representation find only modest lateralization¹ and one study that used narrative stories found highly bilateral results similar to ours². This suggests that right hemisphere areas may respond more strongly to narrative stimuli than to the words and short phrases used in most studies. Still, more research will be needed to determine what roles these left- and right-hemisphere semantic areas have in language comprehension.

Another interesting aspect of these results is that the organization of semantically selective brain areas seems to be highly consistent across individuals. This might suggest that innate anatomical connectivity or cortical cytoarchitecture constrains the organization of high-level semantic representations^{28,29}. It is also possible that this is owing to common life experiences of the subjects, all of whom were raised and educated in Western industrial societies. Future studies that include subjects from more diverse backgrounds will be needed to determine how much of this organizational consistency reflects innate brain structure versus experience.

One limitation of PrAGMATiC as used here is that each area is assumed to be functionally homogeneous. This is a common assumption in the design and analysis of many neuroimaging studies³⁰. However, many cortical maps, including semantic maps in visual cortex¹⁴, seem to contain smoothly changing gradients of representation. It should be possible to modify the PrAGMATiC algorithm to model functional gradients explicitly. This will provide an objective tool for determining whether the semantic maps found here are best described as homogeneous areas or as gradients.

Data-driven approaches are commonplace in studies of human neuroanatomy³¹ and resting state networks^{26,32}, but are only beginning to be used in functional imaging^{14,15}. Our study demonstrates the power and efficiency of data-driven approaches for functional mapping of the human brain. Although our experiment used a simple design in which subjects only listened to stories, the data were rich enough to produce a comprehensive atlas of semantically selective areas. Furthermore, our data-driven framework is quite general. Other properties of language can be mapped (even in this same data set) by using feature spaces that reflect phonemes, syntax and so on. Complex semantic models that incorporate information beyond word co-occurrence can be tested and compared quantitatively. The generalizability of these models can also be tested by using stimuli beyond autobiographical stories. It is sometimes difficult to synthesize the results of data-driven experiments with those from hypothesis-driven experiments, but future methodological and theoretical developments should help to bridge this divide. We expect that the semantic atlas presented here will be useful for many researchers investigating the neurobiological basis of language. We also expect that this atlas can be refined and expanded

by incorporating results from future studies. To facilitate this, we have created a detailed interactive version of the semantic atlas that can be explored online at <http://gallantlab.org/huth2016>.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 8 January 2014; accepted 2 March 2016.

- Binder, J. R., Desai, R. H., Graves, W. W. & Conant, L. L. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb. Cortex* **19**, 2767–2796 (2009).
- Lerner, Y., Honey, C. J., Silbert, L. J. & Hasson, U. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *J. Neurosci.* **31**, 2906–2915 (2011).
- Friederici, A. D., Opitz, B. & von Cramon, D. Y. Segregating semantic and syntactic aspects of processing in the human brain: an fMRI investigation of different word types. *Cereb. Cortex* **10**, 698–705 (2000).
- Noppeney, U. & Price, C. J. Retrieval of abstract semantics. *Neuroimage* **22**, 164–170 (2004).
- Binder, J. R., Westbury, C. F., McKiernan, K. A., Possing, E. T. & Medler, D. A. Distinct brain systems for processing concrete and abstract concepts. *J. Cogn. Neurosci.* **17**, 905–917 (2005).
- Bedny, M., Caramazza, A., Grossman, E., Pascual-Leone, A. & Saxe, R. Concepts are more than percepts: the case of action verbs. *J. Neurosci.* **28**, 11347–11353 (2008).
- Saxe, R. & Kanwisher, N. People thinking about thinking people. The role of the temporo-parietal junction in “theory of mind”. *Neuroimage* **19**, 1835–1842 (2003).
- Caramazza, A. & Shelton, J. R. Domain-specific knowledge systems in the brain the animate-inanimate distinction. *J. Cogn. Neurosci.* **10**, 1–34 (1998).
- Mummery, C. J., Patterson, K., Hodges, J. R. & Price, C. J. Functional neuroanatomy of the semantic system: divisible by what? *J. Cogn. Neurosci.* **10**, 766–777 (1998).
- Just, M. A., Cherkassky, V. L., Aryal, S. & Mitchell, T. M. A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PLoS ONE* **5**, e8622 (2010).
- Warrington, E. K. The selective impairment of semantic memory. *Q. J. Exp. Psychol.* **27**, 635–657 (1975).
- Mitchell, T. M. *et al.* Predicting human brain activity associated with the meanings of nouns. *Science* **320**, 1191–1195 (2008).
- Damasio, H., Grabowski, T. J., Tranel, D., Hichwa, R. D. & Damasio, A. R. A neural basis for lexical retrieval. *Nature* **380**, 499–505 (1996).
- Huth, A. G., Nishimoto, S., Vu, A. T. & Gallant, J. L. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* **76**, 1210–1224 (2012).
- Wehbe, L. *et al.* Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLoS ONE* **9**, e112575 (2014).
- Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M. & Gallant, J. L. Bayesian reconstruction of natural images from human brain activity. *Neuron* **63**, 902–915 (2009).
- Nishimoto, S. *et al.* Reconstructing visual experiences from brain activity evoked by natural movies. *Curr. Biol.* **21**, 1641–1646 (2011).
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **41**, 391–407 (1990).
- Lund, K. & Burgess, C. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behav. Res. Methods Instrum. Comput.* **28**, 203–208 (1996).
- Turney, P. D. & Pantel, P. From frequency to meaning: vector space models of semantics. *J. Artif. Intell. Res.* **37**, 141–188 (2010).
- Caramazza, A. & Mahon, B. Z. The organisation of conceptual knowledge in the brain: the future’s past and some future directions. *Cogn. Neuropsychol.* **23**, 13–38 (2006).
- Huth, A. G., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. PrAGMATiC: a probabilistic and generative model of areas tiling the cortex. Preprint at <http://arxiv.org/abs/1504.03622> (2015).
- Amunts, K., Malikovic, A., Mohlberg, H., Schormann, T. & Zilles, K. Brodmann’s areas 17 and 18 brought into stereotaxic space—where and how variable? *Neuroimage* **11**, 66–84 (2000).
- Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S. & Kanwisher, N. New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *J. Neurophysiol.* **104**, 1177–1194 (2010).
- Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural Comput.* **14**, 1771–1800 (2002).
- Buckner, R. L., Andrews-Hanna, J. R. & Schacter, D. L. The brain’s default network: anatomy, function, and relevance to disease. *Ann. NY Acad. Sci.* **1124**, 1–38 (2008).
- DeWitt, I. & Rauschecker, J. P. Phoneme and word recognition in the auditory ventral stream. *Proc. Natl Acad. Sci. USA* **109**, E505–E514 (2012).
- Riesenhuber, M. Appearance isn’t everything: news on object representation in cortex. *Neuron* **55**, 341–344 (2007).
- Dehaene, S., Cohen, L., Sigman, M. & Vinckier, F. The neural code for written words: a proposal. *Trends Cogn. Sci.* **9**, 335–341 (2005).
- Op de Beeck, H. P., Haushofer, J. & Kanwisher, N. G. Interpreting fMRI data: maps, modules and dimensions. *Nature Rev. Neurosci.* **9**, 123–135 (2008).
- Caspers, S. *et al.* Organization of the human inferior parietal lobule based on receptor architectonics. *Cereb. Cortex* **23**, 615–628 (2013).
- Cohen, A. L. *et al.* Defining functional areas in individual human brains using resting functional connectivity MRI. *Neuroimage* **41**, 45–57 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by grants from the National Science Foundation (NSF; IIS1208203), the National Eye Institute (EY019684), and from the Center for Science of Information (CSol), an NSF Science and Technology Center, under grant agreement CCF-0939370. A.G.H. was also supported by the William Orr Dingwall Neurolinguistics Fellowship. We thank J. Sohl-Dickstein and K. Crane for technical discussions about PrAGMATiC, J. Nguyen for assistance transcribing and aligning stimuli, B. Griffin for segmenting and flattening cortical surfaces, and N. Bilenko, J. Gao, M. Lescroart and A. Nunez-Elizalde for general comments and discussions.

Author Contributions All authors helped conceive and design the experiment. W.A.d.H. and A.G.H. selected and annotated stimuli and collected fMRI data. A.G.H. analysed the data. A.G.H. and T.L.G. designed the PrAGMATiC generative model. A.G.H. and J.L.G. wrote the paper. J.L.G. contributed to all aspects of the project.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.L.G. (gallant@berkeley.edu).

METHODS

MRI data collection. MRI data were collected on a 3T Siemens TIM Trio scanner at the UC Berkeley Brain Imaging Center using a 32-channel Siemens volume coil. Functional scans were collected using gradient echo EPI with repetition time (TR) = 2.0045 s, echo time (TE) = 31 ms, flip angle = 70°, voxel size = $2.24 \times 2.24 \times 4.1$ mm (slice thickness = 3.5 mm with 18% slice gap), matrix size = 100×100 , and field of view = 224×224 mm. Thirty axial slices were prescribed to cover the entire cortex and were scanned in interleaved order. A custom-modified bipolar water excitation radiofrequency (RF) pulse was used to avoid signal from fat. Anatomical data were collected using a T1-weighted multi-echo MP-RAGE sequence on the same 3T scanner.

Subjects. Functional data were collected from five male subjects and two female subjects: S1 (male, age 26), S2 (male, age 32), S3 (female, age 31), S4 (male, age 31), S5 (male, age 26), S6 (female, age 25), and S7 (male, age 30). Two of the subjects were authors (S1: A.G.H.; and S3: W.A.d.H.). All subjects were healthy and had normal hearing. The experimental protocol was approved by the Committee for the Protection of Human Subjects at University of California, Berkeley. Written informed consent was obtained from all subjects. Voxel-wise models were estimated and validated independently for each subject using separate data sets reserved for that purpose. Principal components and PRAGMATIC analyses used leave-one-subject-out cross-validation to verify that the group models accurately predict the data recorded in each individual subject.

Natural story stimuli. The model estimation data set consisted of ten 10- to 15-min stories taken from *The Moth Radio Hour*. In each story, a single speaker tells an autobiographical story in front of a live audience. The ten selected stories cover a wide range of topics and are highly engaging. Each story was played during a separate fMRI scan. The length of each scan was tailored to the story, and included 10 s of silence both before and after the story. These data were collected during two 2-h scanning sessions that were performed on different days. The model validation data set consisted of one 10-min story, also taken from *The Moth Radio Hour*. This story was played twice for each subject (once during each scanning session), and then the two responses were averaged. For story synopses and details of story transcription and preprocessing procedures, see Supplementary Methods.

Stories were played over Sensimetrics S14 in-ear piezoelectric headphones. A Behringer Ultra-Curve Pro hardware parametric equalizer was used to flatten the frequency response of the headphones based on calibration data provided by Sensimetrics. All stimuli were played at 44.1 kHz using the pygame library in Python. All stimuli were normalized to have peak loudness of -1 dB relative to maximum. However, the stories were performed by different speakers and were not uniformly mastered, so some differences in total loudness remain.

Story transcription and preprocessing. Each story was manually transcribed by one listener, and then the transcript was checked by a second listener. Certain sounds (for example, laughter, lip-smacking and breathing) were also marked to improve the accuracy of the automated alignment. The audio of each story was downsampled to 11 kHz and the Penn Phonetics Lab Forced Aligner (P2FA³³) was used to automatically align the audio to the transcript. The forced aligner uses a phonetic hidden Markov model to find the temporal onset and offset of each word and phoneme. The Carnegie Mellon University (CMU) pronouncing dictionary was used to guess the pronunciation of each word. When necessary, words and word fragments that appeared in the transcript but not in the dictionary were manually added. After automatic alignment was complete, Praat³⁴ was used to check and correct each aligned transcript manually. The corrected aligned transcript was then spot-checked for accuracy by a different listener.

Finally, the aligned transcripts were converted into separate word and phoneme representations. The phoneme representation of each story is a list of pairs (p, t) , where p is a phoneme and t is the time from the beginning of the story to the middle of the phoneme (that is, halfway between the start and end of the phoneme) in seconds. Similarly the word representation of each story is a list of pairs (w, t) , where w is a word.

Semantic model construction. To account for response variance caused by the semantic content of the stories, we constructed a 985-dimensional semantic feature space based on word co-occurrence statistics in a large corpus of text^{12,18,19}. First, we constructed a 10,470-word lexicon from the union of the set of all words appearing in the stories and the 10,000 most common words in the large text corpus. We then selected 985 basis words from Wikipedia's *List of 1000 Basic Words* (contrary to the title, this list contained only 985 unique words at the time it was accessed). This basis set was selected because it consists of common words that span a very broad range of topics. The text corpus used to construct this feature space includes the transcripts of 13 *Moth* stories (including the 10 used as stimuli in this experiment), 604 popular books, 2,405,569 Wikipedia pages, and 36,333,459 user comments scraped from reddit.com. In total, the 10,470 words in our lexicon appeared 1,548,774,960 times in this corpus.

Next, we constructed a word co-occurrence matrix, M , with 985 rows and 10,470 columns. Iterating through the text corpus, we added 1 to $M_{i,j}$ each time word j appeared within 15 words of basis word i . A window size of 15 was selected to be large enough to suppress syntactic effects (that is, word order) but no larger. Once the word co-occurrence matrix was complete, we log-transformed the counts, replacing $M_{i,j}$ with $\log(1 + M_{i,j})$. Next, each row of M was z-scored to correct for differences in basis word frequency, and then each column of M was z-scored to correct for word frequency. Each column of M is now a 985-dimensional semantic vector representing one word in the lexicon.

The matrix used for voxel-wise model estimation was then constructed from the stories: for each word–time pair (w, t) in each story we selected the corresponding column of M , creating a new list of semantic vector–time pairs, $(\mathbf{M}_{w,t})$. These vectors were then resampled at times corresponding to the fMRI acquisitions using a 3-lobe Lanczos filter with the cut-off frequency set to the Nyquist frequency of the fMRI acquisition (0.249 Hz).

Voxel-wise model estimation and validation. A linearized finite impulse response (FIR) model^{14,17} consisting of four separate feature spaces was fit to every cortical voxel in each subject's brain. These four feature spaces were word rate (1 feature), phoneme rate (1 feature), phonemes (39 features), and semantics (985 features). The word rate, phoneme rate, and phoneme features were used to account for responses to low-level properties of the stories that could contaminate the semantic model weights (see Supplementary Methods for details of how these low-level models were constructed). A separate linear temporal filter with four delays (1, 2, 3, and 4 time points) was fit for each of these 1,026 features, yielding a total of 4,104 features. This was accomplished by concatenating feature vectors that had been delayed by 1, 2, 3, and 4 time points (2, 4, 6, and 8 s). Thus, in the concatenated feature space one channel represents the word rate 2 s earlier, another 4 s earlier, and so on. Taking the dot product of this concatenated feature space with a set of linear weights is functionally equivalent to convolving the original stimulus vectors with linear temporal kernels that have non-zero entries for 1-, 2-, 3-, and 4-time-point delays.

Before doing regression, we first z-scored each feature channel within each story. This was done to match the features to the fMRI responses, which were also z-scored within each story. However, this had little effect on the learned weights.

The 4,104 weights for each voxel were estimated using L2-regularized linear regression (also known as ridge regression). To keep the scale of the weights consistent and to prevent bias in subsequent analyses, a single value of the regularization coefficient was used for all voxels in all subjects. This regularization coefficient was found by bootstrapping the regression procedure 50 times in each subject. In each bootstrap iteration, 800 time points (20 blocks of 40 consecutive time points each) were removed from the model estimation data set and reserved for testing. Then the model weights were estimated on the remaining 2,937 time points for each of 20 possible regularization coefficients (log spaced between 10 and 1,000). These weights were used to predict responses for the 800 reserved time points, and then the correlation between actual and predicted responses was found. After the bootstrapping was complete, a regularization–performance curve was obtained for each subject by averaging the bootstrap sample correlations first across the 50 samples and then across all voxels. Next, the regularization–performance curves were averaged across the seven subjects and the best overall value of the regularization parameter (183.3) was selected. The best overall regularization parameter value was also the best value in three individual subjects. For the other four subjects the best regularization parameter value was slightly higher (233.6).

To validate the voxel-wise models, estimated semantic feature weights were used to predict responses to a separate story that had not been used for weight estimation. Prediction performance was then estimated as the Pearson correlation between predicted and actual responses for each voxel over the 290 time points in the validation story. Statistical significance was computed by comparing estimated correlations to the null distribution of correlations between two independent Gaussian random vectors of the same length. Resulting P values were corrected for multiple comparisons within each subject using the false discovery rate (FDR) procedure³⁵.

All model fitting and analysis was performed using custom software written in Python, making heavy use of NumPy³⁶, SciPy³⁷, and pycortex³⁸.

Semantic principal components analysis. We used principal components analysis (PCA) to recover a low-dimensional semantic space from the estimated semantic model weights. We first selected only the 10,000 best predicted voxels in each subject according to the average bootstrap correlation (for the selected regularization parameter value) obtained during model estimation. This was done to avoid including noise from poorly modelled voxels. Then we removed temporal information from the voxel-wise model weights by averaging across the four delays for each feature. The weights for the word frequency, phoneme frequency, and phoneme features were then discarded, leaving only the 985 semantic model weights for each voxel. Finally, we applied PCA to these weights, yielding 985 principal components

(PCs). Partial scree plots showing the amount of variance accounted for by each PC are shown in Extended Data Fig. 2. See Supplementary Methods for details.

PrAGMATiC. The PrAGMATiC generative model²² has two components: an arrangement model and an emission model. The arrangement model defines a probability distribution over possible arrangements of the functional areas. This model assumes that the location of each area is defined by a single point called the area centroid. Each centroid is modelled as being joined to nearby centroids by springs. While exact centroid locations can vary from subject to subject, the equilibrium length of each spring is assumed to be consistent across subjects. The probability distribution over possible locations of the centroids is defined using the total potential energy of the spring system. This distribution assigns a high probability to low-energy arrangements of the centroids (that is, where the springs are not stretched much and so store little potential energy) and low probability to high-energy arrangements (where the springs are stretched a lot).

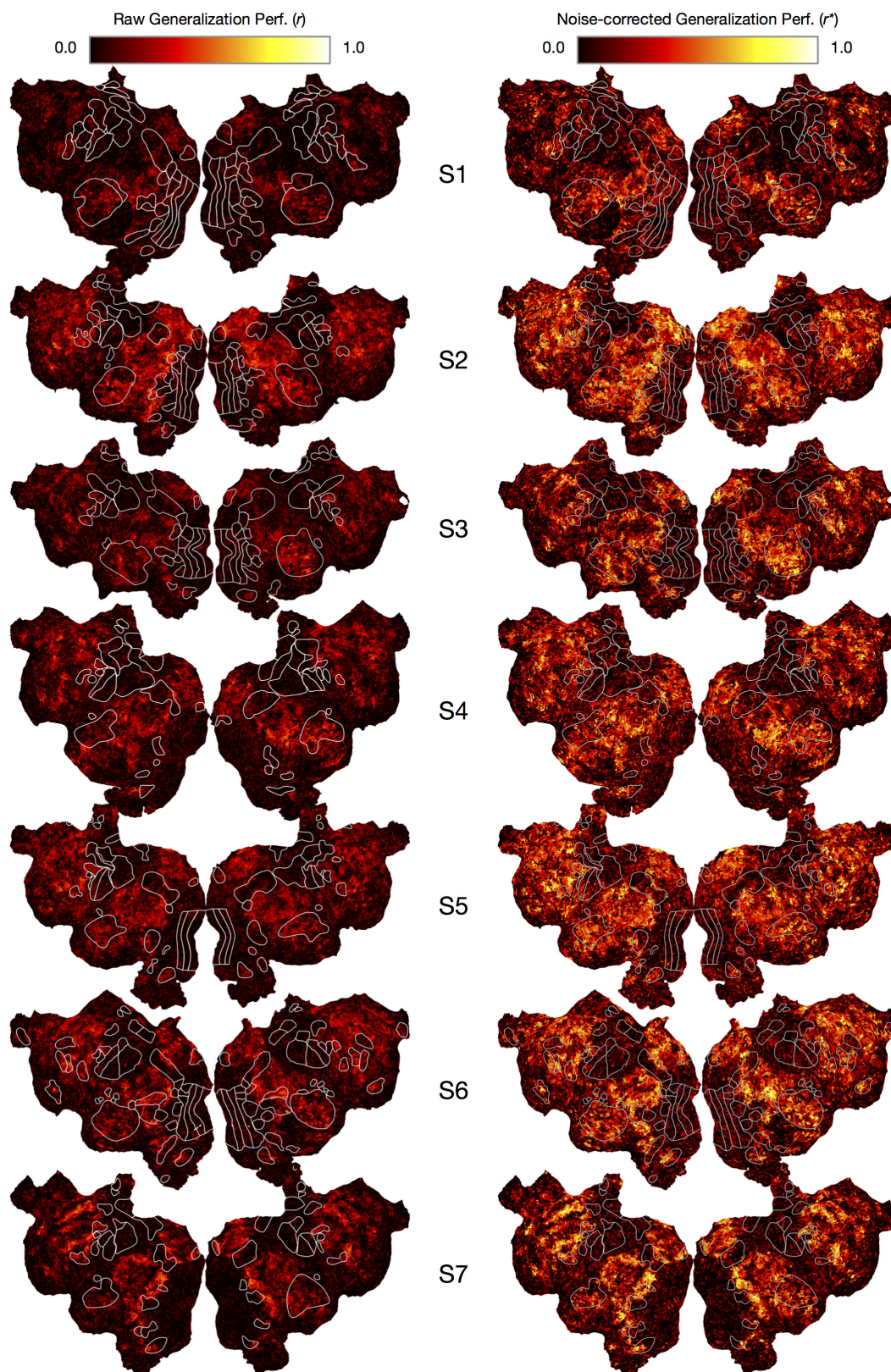
The second component is the emission model, which defines a probability distribution over semantic maps given an arrangement of functional areas. In the emission model each area centroid is assigned a particular semantic value in the four-dimensional common semantic space. This value determines what type of semantic information is represented in that area. To generate a semantic map from any particular arrangement, each point on the cortical surface is first assigned to the closest area centroid (creating a Voronoi diagram). Then the semantic value for each point is sampled from a spherical Gaussian distribution in semantic space, centred on the semantic value of the centroid.

A consequence of modelling semantic maps using a Voronoi diagram is that every point on the cortex must be assigned to an area, while we know that many points on the cortex are not semantically selective. We distinguished between semantically selective and non-selective areas by testing whether the mean semantic voxel-wise model in each area predicted responses significantly better on a held-out story than a baseline model that accounts for responses to phonemes and word rate.

To train the generative model we derived maximum-likelihood estimation (MLE) update rules similar to the Boltzmann learning rule with contrastive divergence²⁵. We used these learning rules to iteratively update the spring lengths and semantic values, maximizing the probability of the observed maps and minimizing the probability of unobserved maps. For details see Supplementary Methods.

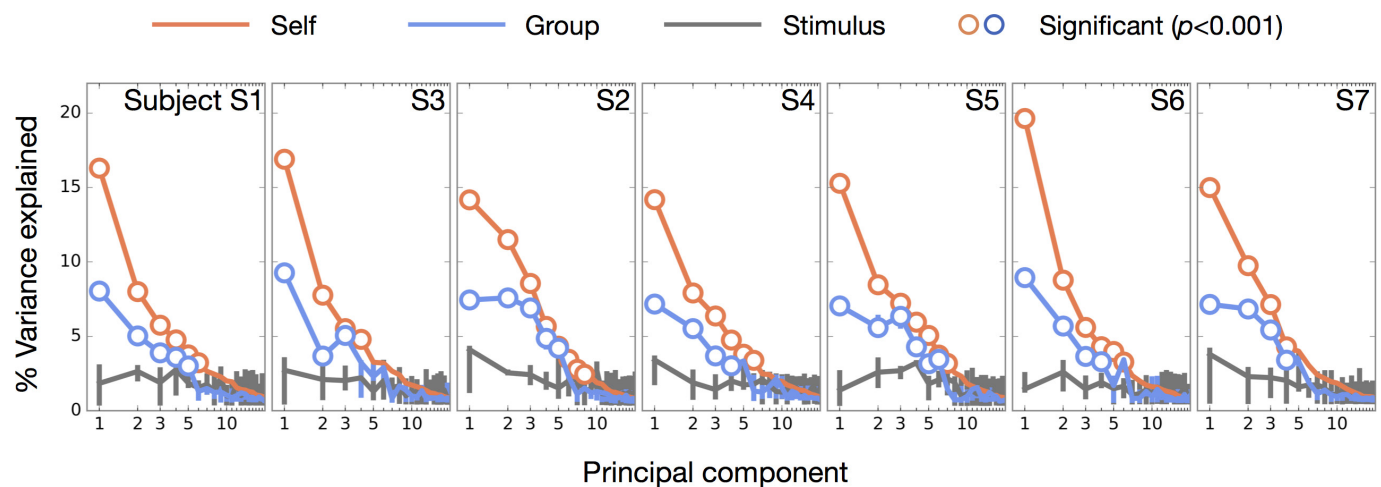
Region of interest abbreviations. Fusiform face area (FFA), occipital face area (OFA), parahippocampal place area (PPA), occipital place area (OPA), retrosplenial cortex (RSC), extrastriate body area (EBA), visual areas (V1-V4, V3A, V3B, V7), lateral occipital visual area (LO), middle temporal visual area (MT+), intraparietal sulcus visual area (IPS), auditory cortex (AC), primary motor and somatosensory areas for feet (M1F, S1F), hands (M1H, S1H), and mouth (M1M, S1M), secondary somatosensory areas for feet (S2F), and hands (S2H), frontal eye fields (FEF), frontal opercular eye movement area (FO), supplementary motor foot area (SMFA), and hand area (SMHA), supplementary eye fields (SEF), Broca's area (BA), superior premotor ventral speech area (sPMv), premotor ventral hand area (PMvh).

33. Yuan, J. & Liberman, M. Speaker identification on the SCOTUS corpus. *Proc. Acoust.* Preprint at <http://www.ling.upenn.edu/~jjahong/publications/c09.pdf> (2008).
34. Boersma, P. & Weenink, D. Praat: doing phonetics by computer (University of Amsterdam, 2014).
35. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **57**, 289–300 (1995).
36. Oliphant, T. E. *Guide to NumPy* (Brigham Young University, 2006).
37. Jones, E., Oliphant, T. E. & Peterson, P. SciPy: Open source scientific tools for Python (SciPy, 2001).
38. Gao, J. S., Huth, A. G., Lescroart, M. D. & Gallant, J. L. Pycortex: an interactive surface visualizer for fMRI. *Front. Neuroinform.* **9**, 23 (2015).



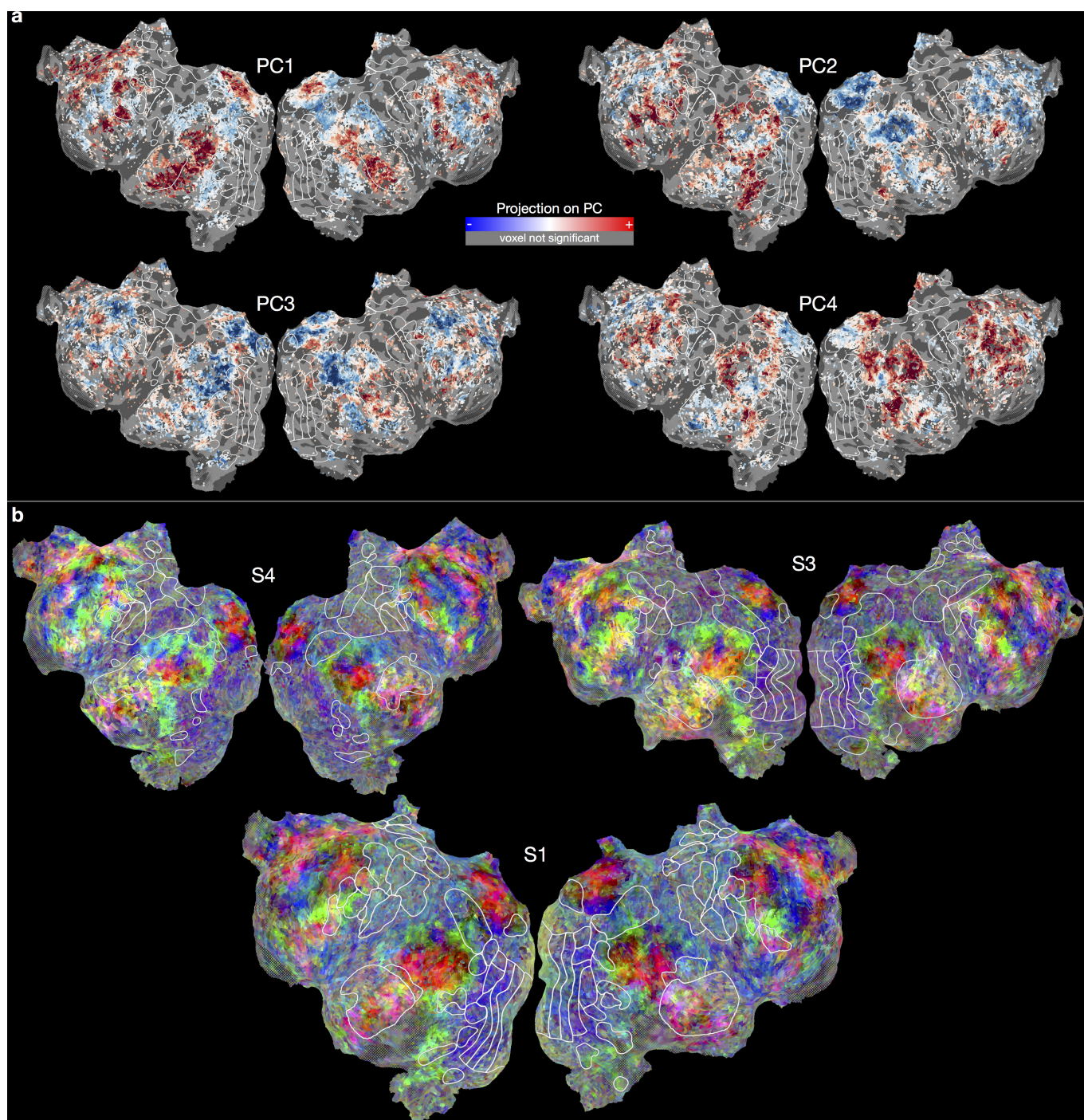
Extended Data Figure 1 | Voxel-wise model prediction performance. Cortical flatmaps showing prediction performance of voxel-wise semantic models for all seven subjects, formatted similarly to Fig. 1c. Models were tested using one 10-min story that was not included during model estimation. Prediction performance was then computed as the correlation between predicted and measured BOLD responses. Left column, raw prediction performance. Note that the colourmap here is scaled 0–1

rather than 0–0.6 as in Fig. 1c to match the scale of the adjusted prediction performance maps. Right column, prediction performance corrected to account for different amounts of noise in the BOLD responses (see Supplementary Methods for details). The voxel-wise semantic models predict BOLD responses in many brain areas, including SPFC, IPFC, LTC, VTC, LPC and MPC. These same regions have been previously identified as the semantic system in the human brain.



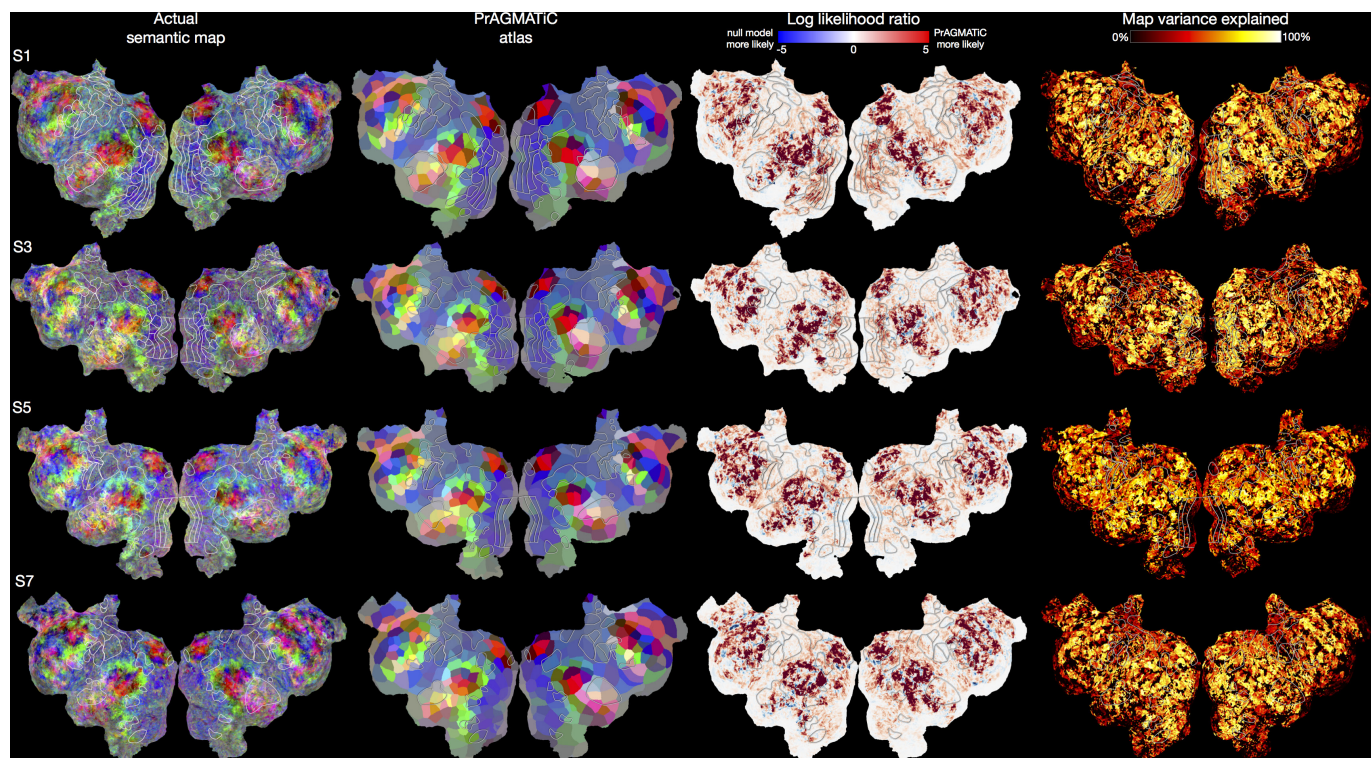
Extended Data Figure 2 | Amount of variance explained by individual subject and group semantic dimensions. Principal components analysis was used to discover the most important semantic dimensions from voxel-wise semantic model weights in each subject. To reduce noise, we used only the 10,000 best voxels in each subject, determined by cross-validation within the model estimation data set. Here we show the amount of variance explained in the semantic model weights by each of the 20 most important principal components (PCs). Orange lines show the amount of variance explained by each subject's own PCs, blue lines show the variance explained by the PCs of combined data from the other six subjects, and grey lines show the variance explained by the PCs of the

stories. (The Gale–Shapley stable marriage algorithm was used to re-order the group and stimulus PCs to maximize their correlation with the subject's PCs.) Error bars indicate 99% confidence intervals. Confidence intervals for the subjects' own PCs and group PCs are very small. Hollow markers indicate subject or group PCs that explain significantly more variance than the corresponding stimulus PCs ($P < 0.001$, bootstrap test). Six PCs explain significantly more variance in one out of seven subjects, five PCs in two subjects, four PCs in three subjects, and three PCs in one subject. Thus, four PCs seem to comprise a semantic space that is common across most individuals.



Extended Data Figure 3 | Separate cortical projections of semantic dimensions 1–4 on subject S2 and combined cortical projections of dimensions 1–3 for subjects S1, S3 and S4. a. Voxel-wise semantic model weights for subject S2 were projected onto each of the common semantic dimensions defined by PCs 1–4. Voxels for which model generalization performance was not significantly greater than zero ($q(\text{FDR}) > 0.05$) are shown in grey. Positive projections are shown in red, negative projections in blue and near-zero projections in white. Voxels with fMRI signal

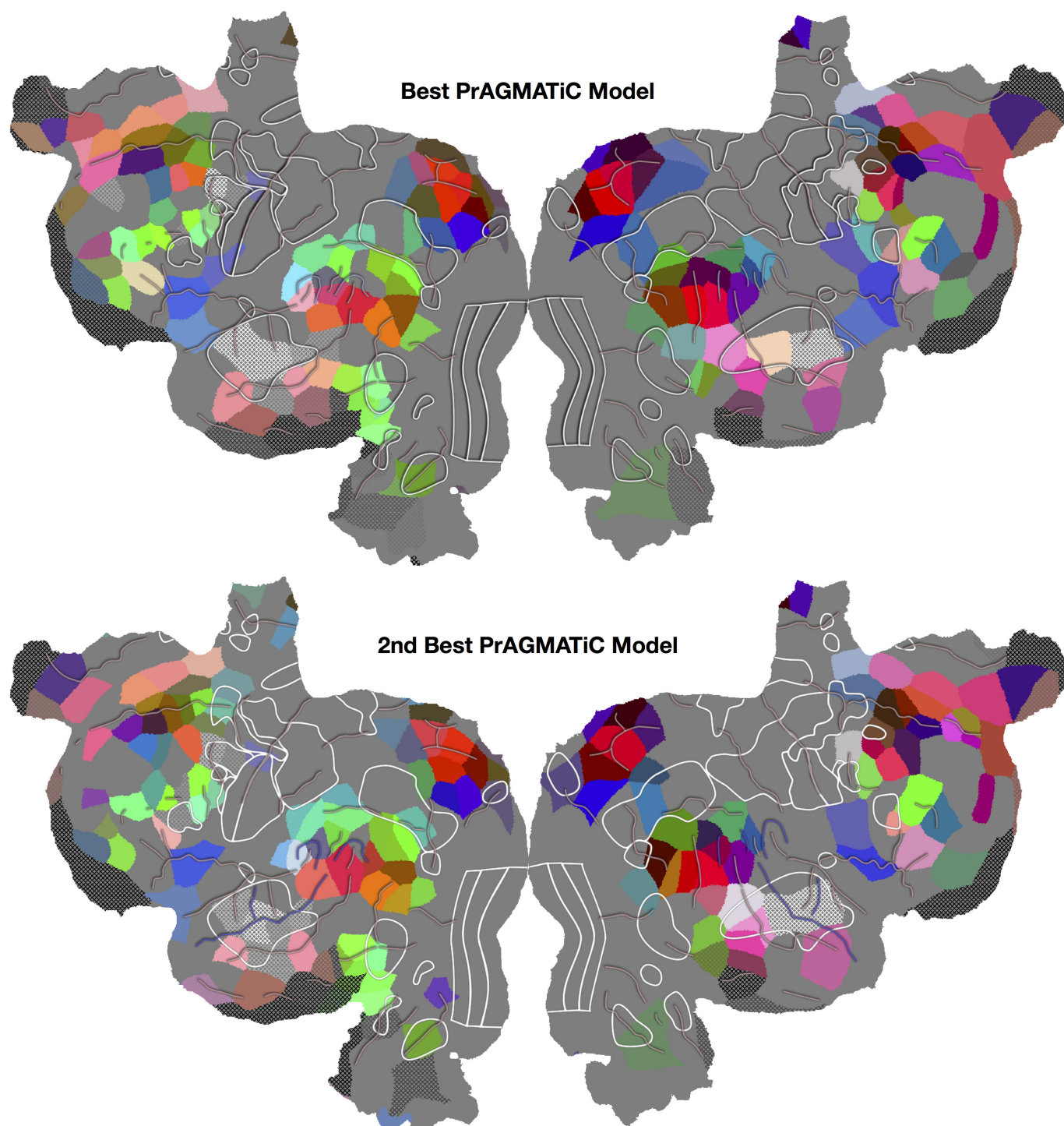
dropout due to field inhomogeneity are shaded with black hatched lines. **b.** Like Fig. 2b, c, this panel shows the result of projecting voxel-wise models onto the first three common semantic dimensions, and then colouring each voxel using an RGB colourmap. The red colour component corresponds to the projection on the first PC, the green component to the second, and the blue component to the third. Semantic information seems to be represented in complex patterns distributed across the semantic system and the patterns seem to be largely conserved across individuals.



Extended Data Figure 4 | PrAGMATiC atlas likelihood maps.

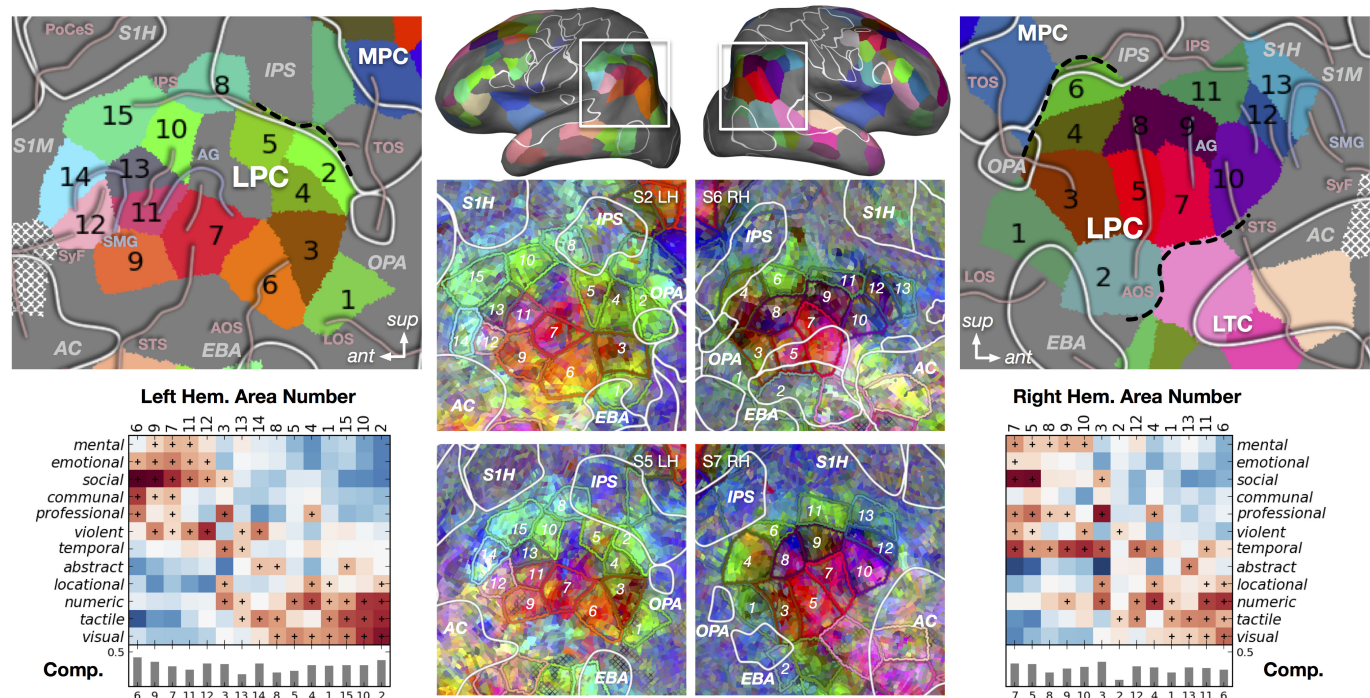
Comparison of actual semantic maps (Fig. 2, Extended Data Fig. 3) to the maps generated from the PrAGMATiC atlas (Fig. 3). PrAGMATiC atlases for the left and right hemispheres were fit using data from all seven subjects. The left hemisphere atlas has 192 total areas and the right hemisphere has 128 (including non-semantic areas). Here we show the actual semantic maps for four subjects (first column), the PrAGMATiC atlas on each subject's cortical surface (second column), the log likelihood ratio of the actual semantic map under the PrAGMATiC atlas versus a null model (third column), and the fraction of variance in the semantic map that the PrAGMATiC atlas explains for each location on the cortical surface (fourth column). The likelihood ratio maps show that most

areas where there are large semantic model weights (that is, the semantic system) are much better explained by PrAGMATiC than by a null model and thus appear red, while areas where the weights are small (that is, somatomotor cortex, visual cortex, and so on) are about equally well explained by both PrAGMATiC and the null model and thus appear white. Variance explained was computed by subtracting the PrAGMATiC atlas from the actual semantic map (in the space of the four group semantic dimensions), squaring and summing the residuals and then dividing by the sum of squares in the actual map. The variance explained maps show that the PrAGMATiC atlas captures a large fraction of the variance in the semantic maps (37–47% in total).



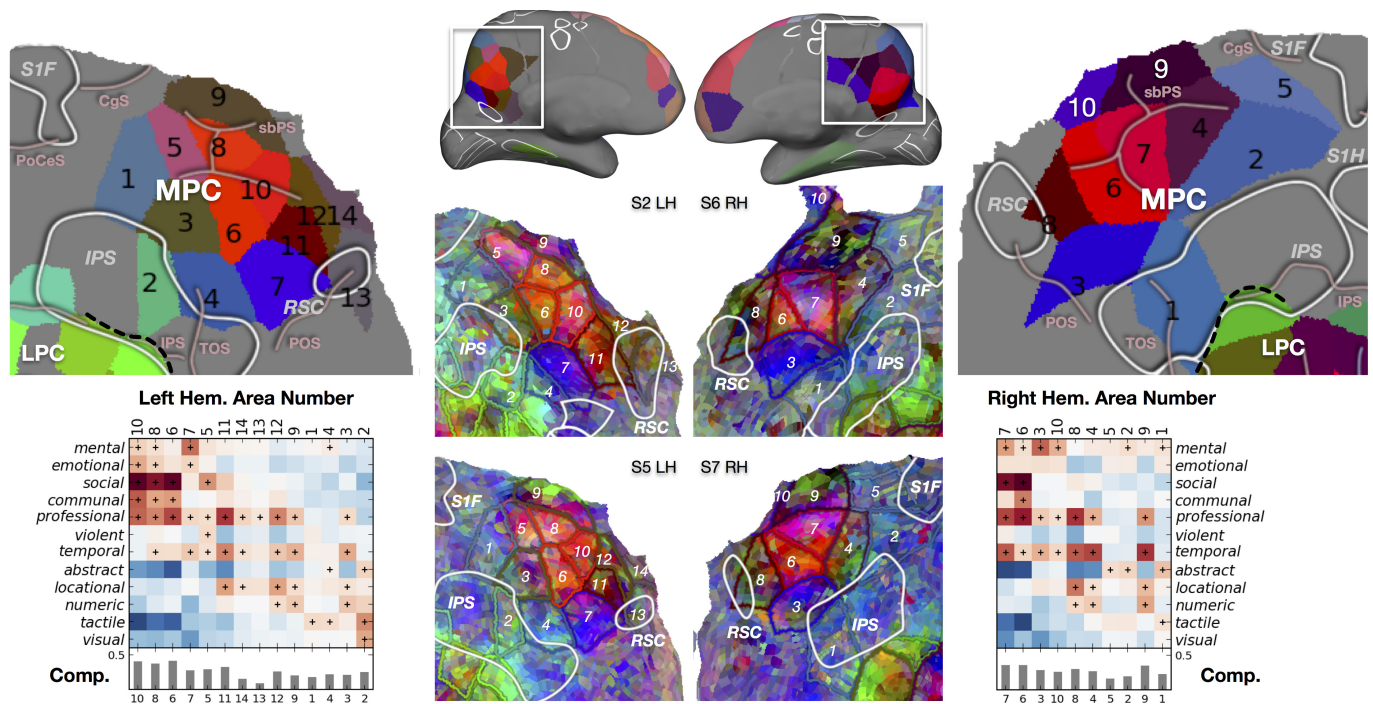
Extended Data Figure 5 | Comparison of PrAGMATiC models fit with different initial conditions. As with many clustering algorithms, PrAGMATiC optimizes a non-convex objective function and so can find many potential locally optimal solutions. To reduce the effect of non-convexity on our results, we re-fit the model ten times (each time with a different random initialization), and then selected the model fit that yielded the best likelihood (that is, performance on the training set) as the PrAGMATiC atlas (Fig. 3). Here we show the PrAGMATiC atlas (top) and the second best model out of the ten that were estimated (bottom). The parcellations given by these two models are very similar. However, there are a few differences, which illustrate uncertainty in the model.

Some of these differences are due to statistical thresholding: a few areas that were found to be significantly semantically selective in the best model are missing in the alternative model (see left medial prefrontal cortex), and some significant areas in the alternate model are missing from the best model (left ventral occipital cortex). Other differences suggest alternative parcellations for a few regions, where, for example, the same region of cortex is parcellated into three areas in the best model and four areas in the alternative model. Yet it is clear that none of the differences between these two models are sufficient to change any of the interpretations given in the main text.



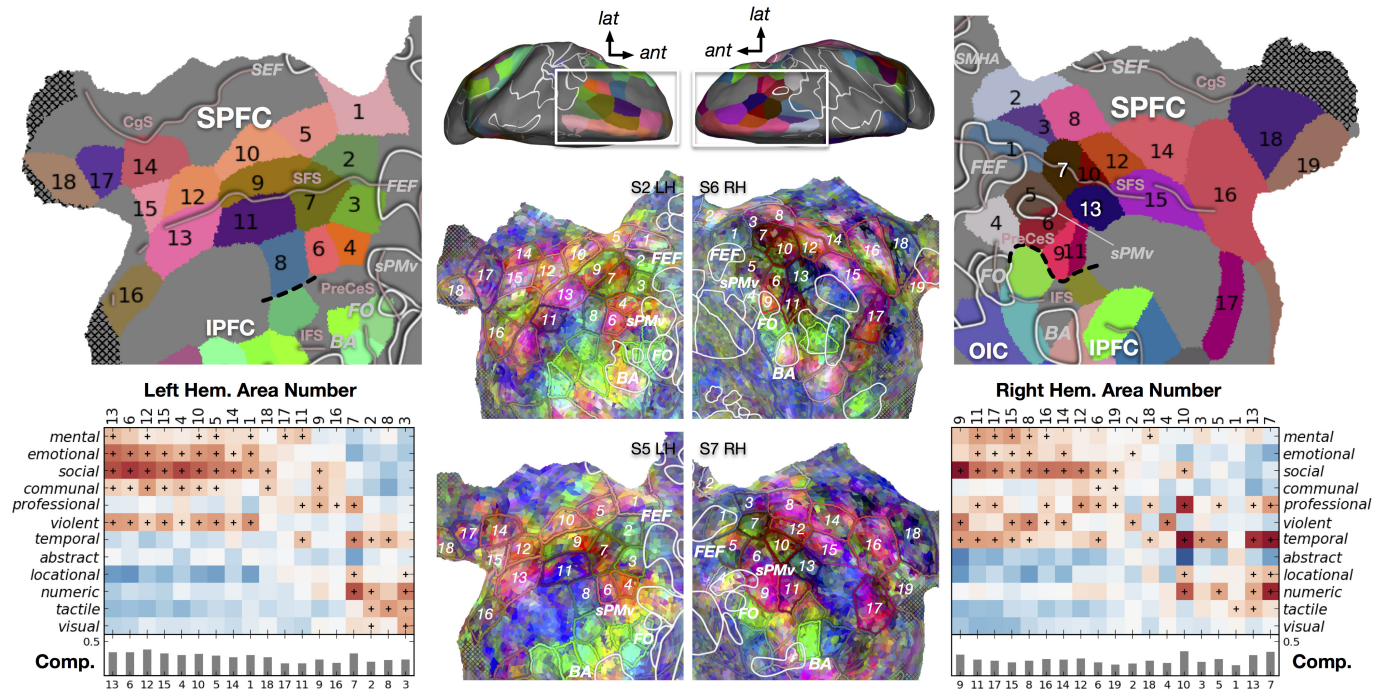
Extended Data Figure 6 | Semantic atlas for the LPC. The PrAGMATiC atlas divides the LPC into 15 areas in the left hemisphere and 13 areas in the right. Here we show the atlas for each hemisphere (top left and right), three-dimensional brains indicating the location of the LPC (top middle), individual maps for two subjects in each hemisphere (bottom middle), and the average predicted response of each area to the 12 semantic categories identified earlier (responses consistently greater than zero across subjects

are marked with a plus) (bottom left and right). Bars show how completely this 12-category interpretation captures the average semantic model in each area. The LPC appears to be organized around the angular gyrus (AG), with a core that is selective for social, emotional and mental concepts (L6, 7, 9, 11; R5, 7) and a periphery that is selective for visual, tactile and numeric concepts (L2, 4, 5, 8, 10, 15; R6, 11).



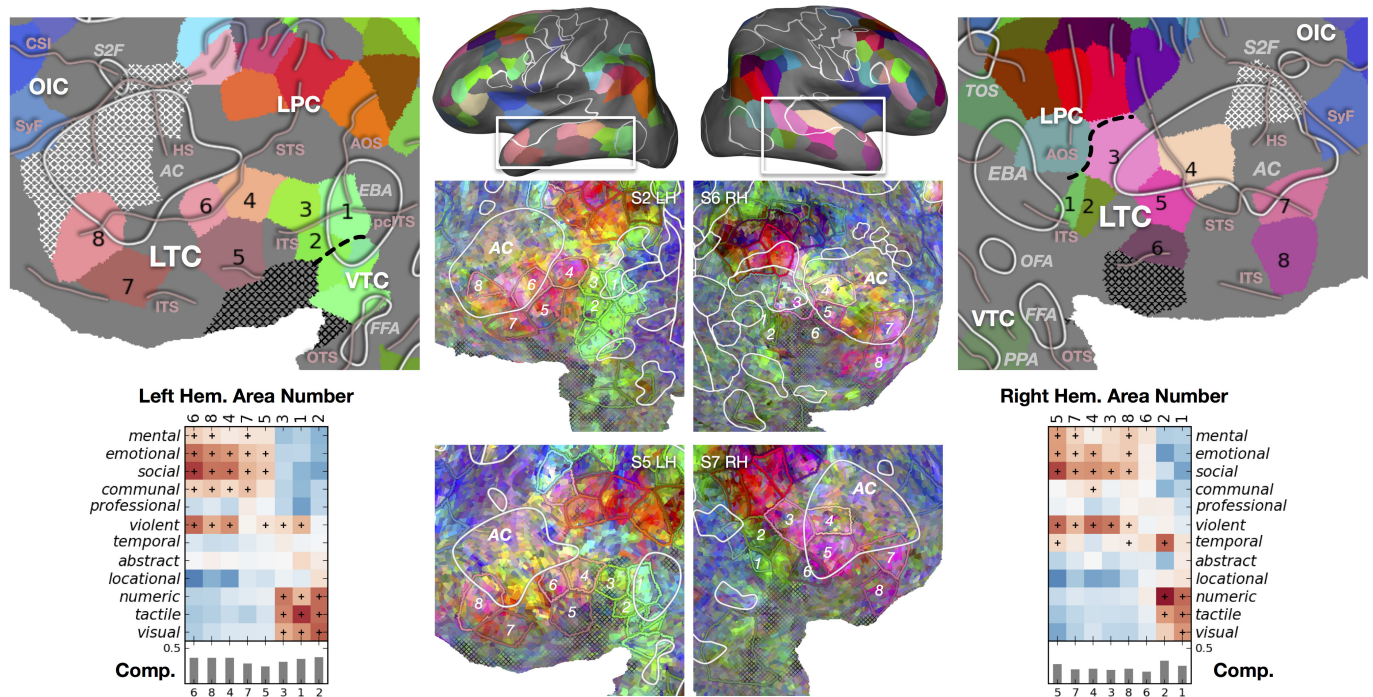
Extended Data Figure 7 | Semantic atlas for the MPC. The PrAGMATiC atlas divides the MPC into 14 areas in the left hemisphere and 10 areas in the right. Here we show the atlas for each hemisphere (top left and right), three-dimensional brains indicating the location of the MPC (top middle), individual maps for two subjects in each hemisphere (bottom middle), and the average predicted response of each area to the 12 semantic categories identified earlier (responses consistently greater than zero across subjects are marked with a plus) (bottom left and right). Bars show how completely the 12-category interpretation captures the

average semantic model in each area. Like the LPC, the MPC appears to be organized around a core group of areas that are selective for social and mental concepts (L6, 8, 10; R6, 7). Dorsolateral MPC areas (L2, 4; R1) are selective for visual and tactile concepts. Anterior dorsal areas (L5, 9; R4, 9) are selective for temporal concepts. Ventral areas (L11, 12, 14; R8) are selective for professional, temporal and locational concepts. Just above the retrosplenial cortex one distinct area in each hemisphere is selective for mental, professional and temporal concepts (L7; R3). Overall, the right MPC responds more than the left MPC to mental concepts.



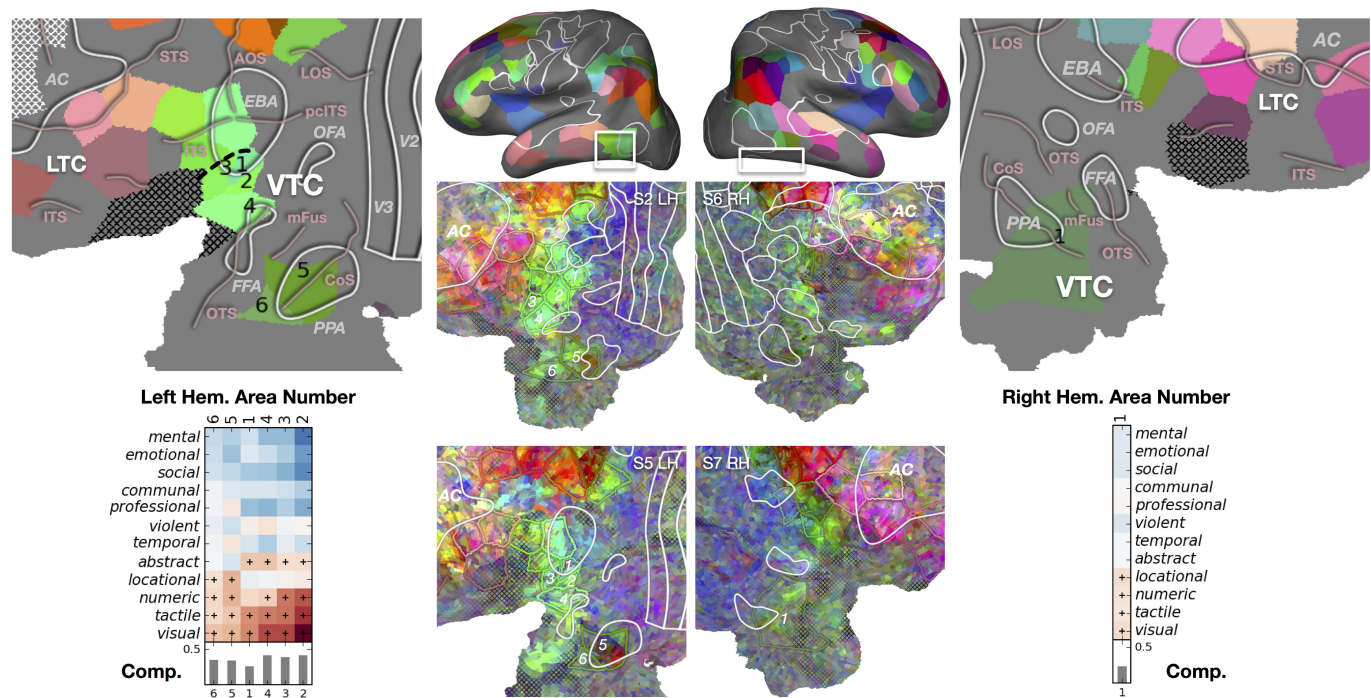
Extended Data Figure 8 | Semantic atlas for the SPFC. The PrAGMATiC atlas divides the SPFC into 18 areas in the left hemisphere and 19 areas in the right. Here we show the atlas for each hemisphere (top left and right), three-dimensional brains indicating the location of the SPFC (top middle), individual maps for two subjects in each hemisphere (bottom middle), and the average response of each area in the atlas to the 12 semantic categories identified earlier (responses consistently greater than zero across subjects are marked with a plus) (bottom left and right). Bars show how completely

the 12-category interpretation captures the average semantic model in each area. The organization in the SPFC seems to follow the long rostro-caudal sulci and gyri of the dorsal frontal lobe. Posterior-lateral SPFC areas (L4, 6; R6, 9, 11) are selective for social, emotional, communal and violent concepts. Posterior superior frontal sulcus areas (L2, 3, 7, 8; R1, 5, 7) are selective for visual, tactile and numeric concepts. The superior frontal gyrus contains a long strip of areas (L1, 5, 10, 12–15; R8, 12, 14–16) selective for social, emotional, communal and violent concepts.



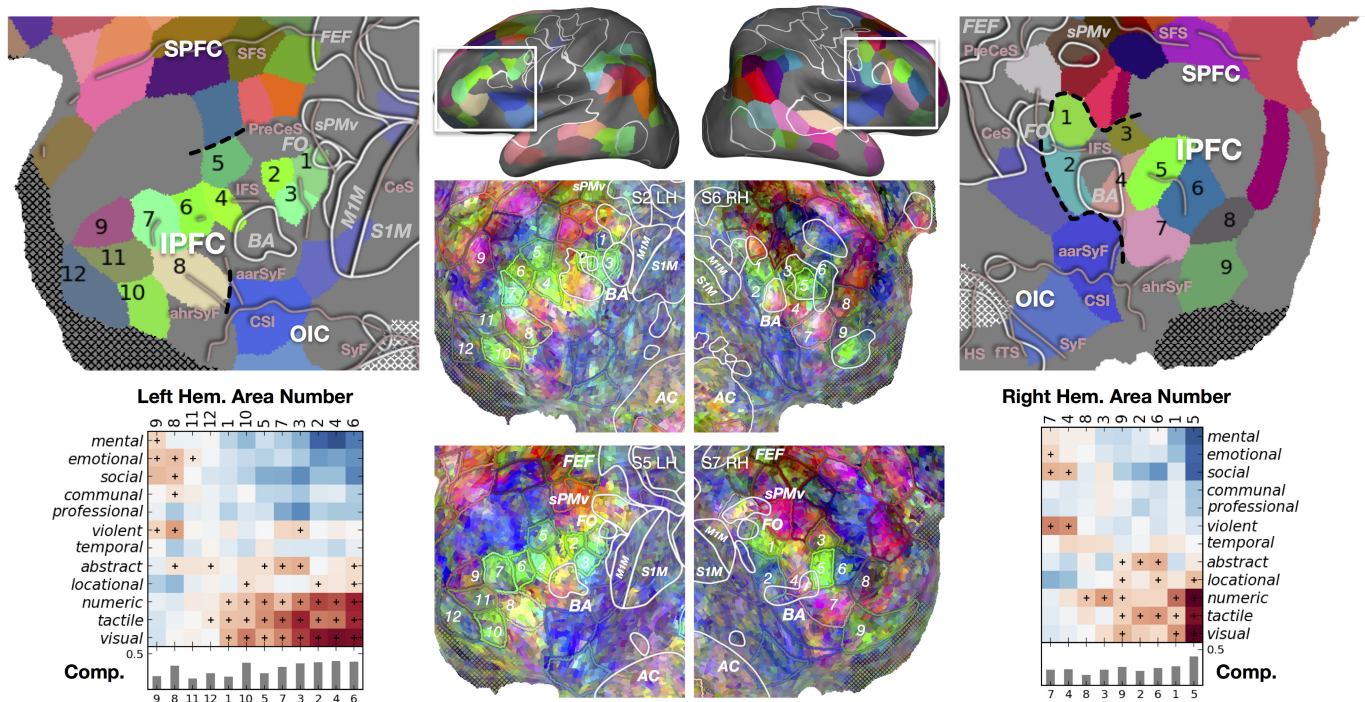
Extended Data Figure 9 | Semantic atlas for the LTC. The PrAGMATiC atlas divides the LTC into eight areas in both the left and right hemispheres. Here we show the atlas for each hemisphere (top left and right), three-dimensional brains indicating the location of the LTC (top middle), individual maps for two subjects in each hemisphere (bottom middle), and the average response of each area in the atlas to the 12

semantic categories identified earlier (responses consistently greater than zero across subjects are marked with a plus) (bottom left and right). Bars show how completely the 12-category interpretation captures the average semantic model in each area. Anterior LTC areas (L4–8; R3–8) are selective for social, emotional, mental and violent concepts. Posterior LTC areas (L1–3; R1–2) are selective for numeric, tactile and visual concepts.



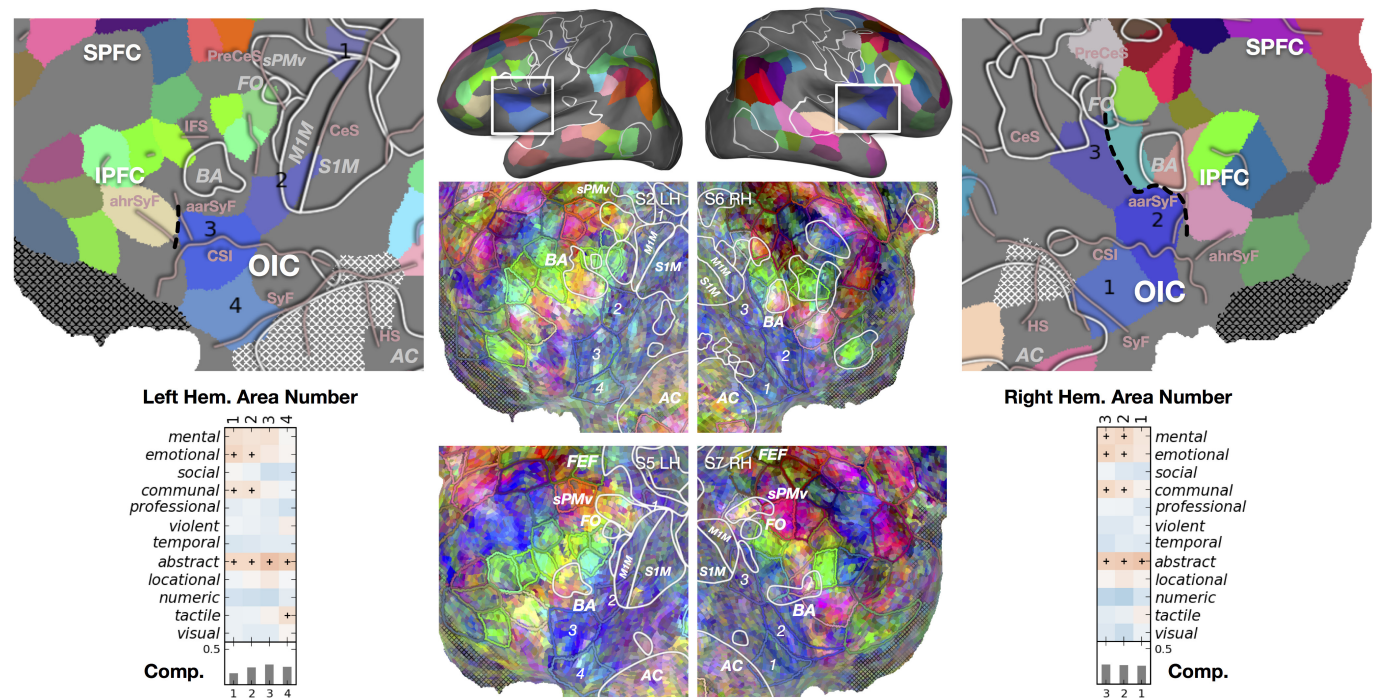
Extended Data Figure 10 | Semantic atlas for the VTC. The PrAGMATiC atlas divides the VTC into six areas in the left hemisphere and one area in the right. Here we show the atlas for each hemisphere (top left and right), three-dimensional brains indicating the location of the VTC (top middle), individual maps for two subjects in each hemisphere (bottom middle), and the average response of each area in the atlas to the 12 semantic categories identified earlier (responses consistently greater

than zero across subjects are marked with a plus) (bottom left and right). Bars show how completely the 12-category interpretation captures the average semantic model in each area. The VTC is relatively homogeneous: all areas are selective for numeric, tactile and visual concepts. Left VTC areas close to the parahippocampal place area (PPA) are also selective for locational concepts (L5–6).



Extended Data Figure 11 | Semantic atlas for the IPFC. The PrAGMATiC atlas divides the IPFC into 12 areas in the left hemisphere and 9 areas in the right. Here we show the atlas for each hemisphere (top left and right), three-dimensional brains indicating the location of the IPFC (top middle), individual maps for two subjects in each hemisphere (bottom middle), and the average response of each area in the atlas to the 12 semantic categories identified earlier (responses consistently greater than zero across subjects are marked with a plus) (bottom left

and right). Bars show how completely the 12-category interpretation captures the average semantic model in each area. Posterior IPFC areas in the precentral sulcus (L1–3; R1, 2) are selective for visual, tactile and numeric concepts. Areas on the inferior frontal gyrus (L8; R4, 7) are selective for social and violent concepts. Areas in the inferior frontal sulcus and anterior middle frontal gyrus (L4–7; R5–6) are selective for visual, tactile and numeric concepts. Areas in the orbitofrontal sulci (L10; R9) are also selective for visual, tactile, numeric and locational concepts.



Extended Data Figure 12 | Semantic atlas for the opercular and insular cortex. The PrAGMATiC atlas divides the opercular and insular cortex (OIC) into four areas in the left hemisphere and three areas in the right. Here we show the atlas for each hemisphere (top left and right), three-dimensional brains indicating the location of the OIC (top middle), individual maps for two subjects in each hemisphere (bottom middle), and the average response of each area in the atlas to the 12 semantic

categories identified earlier (responses consistently greater than zero across subjects are marked with a plus) (bottom left and right). Bars show how completely the 12-category interpretation captures the average semantic model in each area. These areas are homogeneously selective for abstract concepts, with more posterior and superior areas also responding to emotional, communal and mental concepts.

Robust neuronal dynamics in premotor cortex during motor planning

Nuo Li^{1*}, Kayvon Daie^{1*}, Karel Svoboda¹ & Shaul Druckmann¹

Neural activity maintains representations that bridge past and future events, often over many seconds. Network models can produce persistent and ramping activity, but the positive feedback that is critical for these slow dynamics can cause sensitivity to perturbations. Here we use electrophysiology and optogenetic perturbations in the mouse premotor cortex to probe the robustness of persistent neural representations during motor planning. We show that preparatory activity is remarkably robust to large-scale unilateral silencing: detailed neural dynamics that drive specific future movements were quickly and selectively restored by the network. Selectivity did not recover after bilateral silencing of the premotor cortex. Perturbations to one hemisphere are thus corrected by information from the other hemisphere. Corpus callosum bisections demonstrated that premotor cortex hemispheres can maintain preparatory activity independently. Redundancy across selectively coupled modules, as we observed in the premotor cortex, is a hallmark of robust control systems. Network models incorporating these principles show robustness that is consistent with data.

Neurons in frontal and parietal cortex show slow dynamics, including persistent and ramping activity, related to motor planning^{1–4}, action timing^{5,6}, working memory^{7–10} and decision making^{11–13}. Neurons have intrinsic time constants on the order of ten milliseconds¹⁴. Slow dynamics over seconds are presumably an emergent property of neural circuits, probably involving feedback drive¹⁵ (but also see refs 16, 17).

Network models can produce persistent and ramping activity, including integrators^{15,18–21} and trained recurrent networks^{22–24}. The amplification that prolongs the model network response may cause fragility to perturbations of activity²⁵. By contrast, biological systems are typically robust to internal and external perturbations^{26,27}.

Controlled transient perturbations can probe the mechanisms underlying the dynamics in neural networks^{13,25,28,29}. Deviations from normal activity patterns are related to network structure. For example, attractor-like models predict recovery of the attractor state with altered dynamics, whereas chaotic systems diverge over time²⁵. Comparison of perturbed dynamics and behaviour can reveal which elements of the original dynamics are necessary.

We measured behavioural and neural responses after transiently silencing parts of the mouse premotor cortex (anterior lateral motor cortex, ALM). ALM neurons in both hemispheres, which are coupled via callosal axons, exhibit persistent preparatory activity that predicts specific movement directions, seconds before the movement^{3,30}. We report that preparatory activity is robust to unilateral perturbations. Theoretical analyses suggest that premotor networks are organized into redundant modules.

Preparatory activity in ALM

Mice performed pole location discrimination with their whiskers^{3,30} (Fig. 1a). During a subsequent delay epoch (1.3–1.7 s), mice planned the upcoming response. An auditory ‘go’ cue (0.1 s) signalled the beginning of the response epoch, and mice reported pole position by licking one of two ports (posterior to lick right; anterior to lick left).

ALM is involved in planning directional licking^{3,30,31}. We recorded single units from the left ALM ($n = 1,012$ units from 12 mice; Methods) (Fig. 1b). Most ALM pyramidal neurons distinguished trial types (634 out of 890, $P < 0.05$, t -test; sample epoch, 176 out of 890; delay

epoch, 337 out of 890; response epoch, 493 out of 890; Methods) (Extended Data Fig. 1). Selectivity was defined as the spike rate difference between ‘lick left’ and ‘lick right’ trials. Individual ALM neurons exhibited diverse patterns of activity during different task epochs, including persistent activity and ramping activity during the delay epoch, similar to activity seen across the frontal cortex^{1,2,4,5,8–10,32,33}, parietal cortex^{6,12} and subcortical brain areas³⁴.

Preparatory activity after unilateral silencing

Models of persistent and ramping activity^{5,18,22–24,35–37} do not recover after transiently silencing comparably sized subsets of neurons (Fig. 1c and Extended Data Fig. 1f–i). We transiently silenced preparatory activity³ (Fig. 1a) (‘photoinhibition’; Extended Data Fig. 2 and Methods). The standard photostimulus was one laser spot³, silencing 58% of one ALM hemisphere ($>80\%$ reduction of activity, Methods) (Fig. 1b). Transient (duration, 0.5 s) unilateral photoinhibition of ALM up to the go cue (late delay) caused an ipsilateral response bias ($n = 5$ mice, $P < 0.01$, two-tailed t -test; Fig. 1d), similar in magnitude to photoinhibition over the entire delay epoch^{3,30} (Extended Data Fig. 3). By contrast, photoinhibition ending at least 0.3 s before the go cue produced minimal behavioural effects (middle delay, early delay; $P > 0.1$, two-tailed t -test).

ALM activity was abolished during photoinhibition (Fig. 1e) ($n = 6$ mice). After photoinhibition offset preparatory activity recovered. ALM neurons that normally exhibited ramping activity during the delay epoch accelerated their ramping after photoinhibition so that activity ‘caught up’ to reach the same level as in unperturbed trials (Fig. 1e, neurons 1 and 2). Recovery was not due to non-specific overshoots in spike rate after photoinhibition (that is, ‘rebound’). First, we used photostimuli optimized to minimize rebound³ (Extended Data Fig. 2c). Second, selectivity also recovered, so that activity reached the appropriate spike rate for each trial type. Finally, neurons that normally did not exhibit increasing ramps during the delay epoch also recovered their activity (Fig. 1e, neuron 3; Extended Data Fig. 4). Within 400 ms of photoinhibition, spike rates became indistinguishable from the unperturbed condition in 90% of the neurons; only 10% of neurons retained a sustained change in spike rate (Fig. 1f). ALM neurons recovered $>80\%$ of their selectivity relative to the unperturbed trials within 514 ms of photoinhibition (Fig. 1g).

¹Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, Virginia 20147, USA.

*These authors contributed equally to this work.

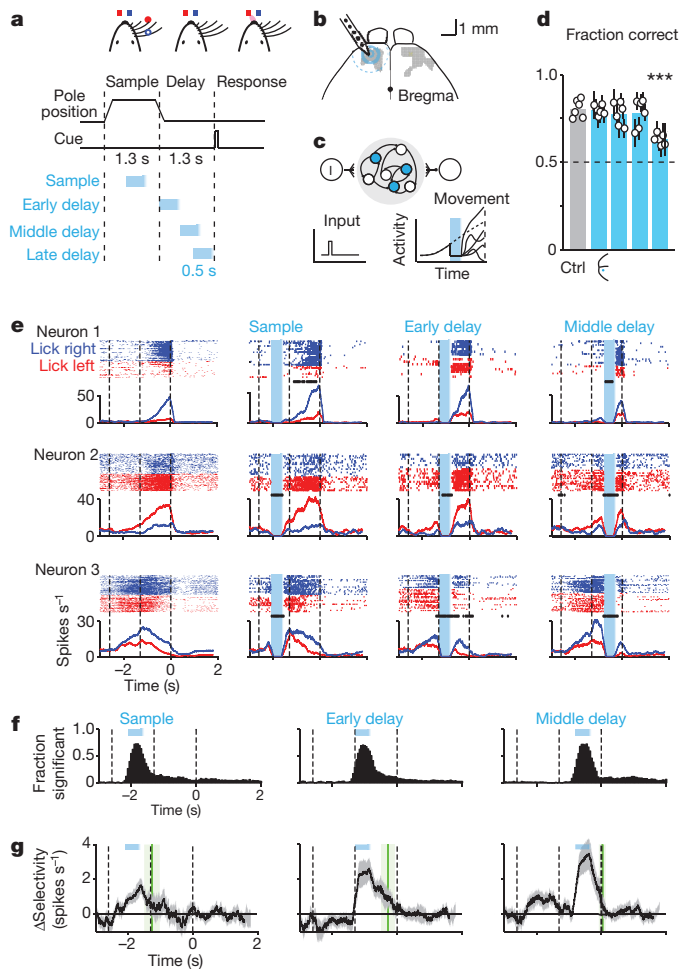


Figure 1 | ALM preparatory activity is robust to photoinhibition.

a, Mice discriminate pole location during the sample epoch and respond 'lick right' or 'lick left' after a delay. Cyan, photoinhibition. **b**, Grey, ALM; area that produced behavioural effects with photoinhibition throughout the delay epoch (Methods; Allen Mouse Brain Atlas (<http://mouse.brain-map.org/static/atlas>). Cyan, contours of photoinhibition (small, 90% reduction in activity; medium, 80%; large/dashed, 50%). **c**, Schematic network models and responses to transient photoinhibition of subsets of neurons (cyan). Dashed line, unperturbed activity trajectory; solid line, perturbed activity trajectories. **d**, Behavioural performance (see timing in **a**). Bar, mean. Symbols, individual mice (mean \pm s.e.m., bootstrap). *** $P < 0.001$, two-tailed t -test against control. **e**, Example neurons. Top, spike raster. Bottom, peristimulus time histogram (PSTH), averaged over 200 ms. Lick-right (blue) and lick-left (red) trials, grouped by instructed movement. Dashed lines, behavioural epochs. Cyan, photoinhibition. Black ticks above PSTH, significant spike rate change ($P < 0.01$, two-tailed t -test). **f**, Fraction of neurons with significant spike rate change ($n = 168$, 168 and 175). Cyan, photoinhibition. **g**, Δ Selectivity from control (mean \pm s.e.m. across neurons, bootstrap; selective neurons tested for > 3 trials in all conditions, $n = 55$). Green lines, recovery to 80% of control (mean \pm s.e.m. bootstrap). Sample, 373 ± 260 ms; early delay, 510 ± 218 ms; middle delay, 327 ± 112 ms.

In separate experiments, we transiently (500 ms) photostimulated a subset of layer 5 pyramidal neurons³⁰ (Extended Data Fig. 5). After photostimulus offset, ALM activity and selectivity recovered with a time-course that was similar to recovery after photoinhibition. Thus, ALM premotor activity is robust to large perturbations of activity.

Preparatory activity after bilateral silencing

Perturbed ALM probably inherits preparatory activity from a connected area. ALM is bilaterally connected through the corpus callosum, and preparatory activity is found in both hemispheres³⁰. We

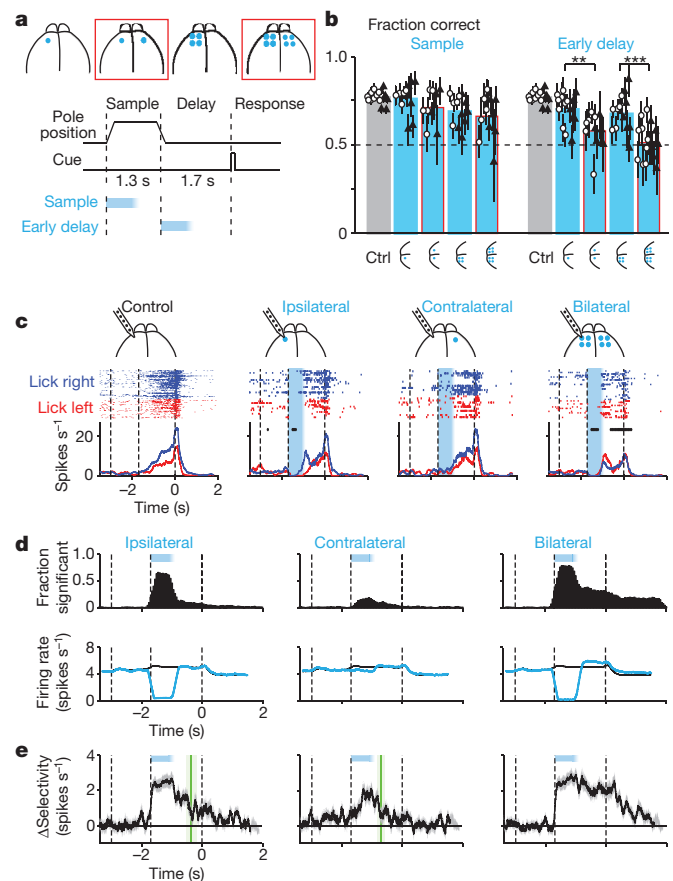


Figure 2 | Bilateral photoinhibition disrupts preparatory activity.

a, Unilateral and bilateral (red) photoinhibition. **b**, Behavioural performance. Bar, mean. Symbols, individual mice (mean \pm s.e.m., bootstrap). Open circle, photoinhibition duration, 800 ms; solid triangle, 1,300 ms. * $P < 0.01$, *** $P < 0.001$, two-tailed t -test against control. **c**, Example ALM neuron. Cyan, photoinhibition. **d**, Fraction of neurons with significant spike rate change ($n = 276$, 283 and 332). Bottom, average spike rate across the population (black, control; cyan, photoinhibition). **e**, Average change in population selectivity from control ($n = 143$). Same as Fig. 1g. Selectivity recovery: ipsilateral, 538 ± 178 ms; contralateral, 192 ± 114 ms; bilateral, no recovery.

tested for coupling between hemispheres by silencing ALM activity either unilaterally or bilaterally ($n = 13$ mice) (Fig. 2a; Methods), using four protocols: (1) unilateral photoinhibition with one laser spot (left or right hemisphere); (2) bilateral photoinhibition using one spot on each side; (3) unilateral photoinhibition using a grid of four spots (1-mm spacing), silencing all of ALM and surrounding regions (Fig. 1b); and (4) bilateral photoinhibition using four spots on each side. Photoinhibition (duration 0.8 s or 1.3 s) was deployed during either the sample or early in the delay epoch, ending at least 0.4 s before the response cue (Fig. 2a).

Behavioural performance was only slightly affected after unilateral photoinhibition with a single spot during the early delay epoch (Fig. 2b, 70.3% correct, $P = 0.009$, two-tailed t -test against control); unilateral photoinhibition with four spots had a small additional effect (67.7%, $P = 0.003$). By contrast, using only two spots across both hemispheres caused performance to degrade severely (Fig. 2b, 58.0%, $P < 0.001$; difference from four spot unilateral: $P < 0.05$, two-tailed t -test); four spots bilaterally further reduced performance to near chance level (four laser spots: 51.4%, $P < 0.001$). This implies that the larger effects of bilateral photoinhibition were not simply due to the strength of photoinhibition. Bilateral photoinhibition biased movements inconsistently across mice and sessions (Extended Data Fig. 6a); we use this feature later to explore the relationship between ALM population dynamics and movement.

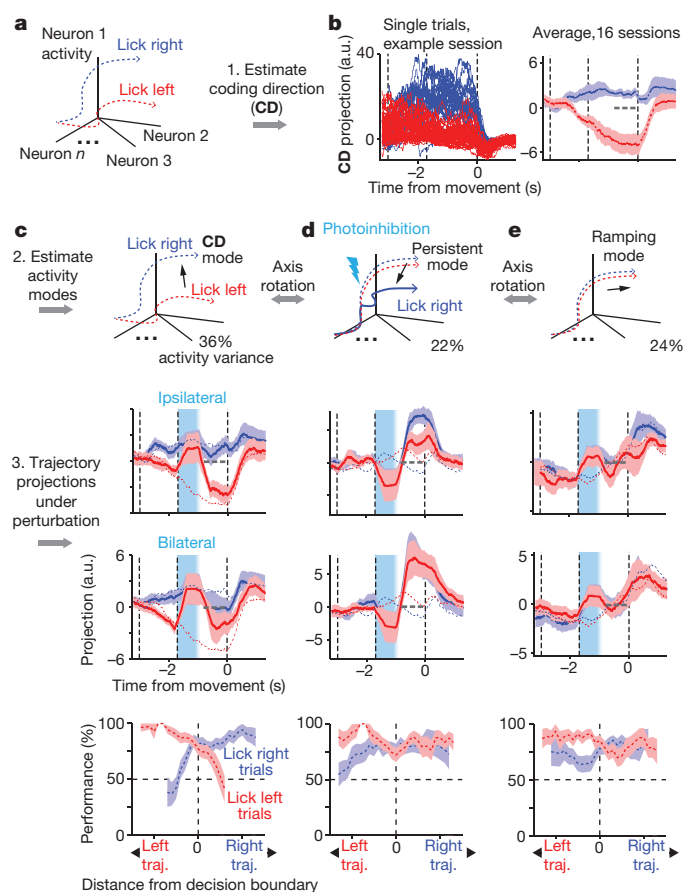


Figure 3 | Preparatory activity preferentially recovers along coding dimension in activity space. **a**, Schematic, movement-specific trajectories in activity space. **b**, Left, activity on correct lick-right (blue) and lick-left (red) trials projected onto the coding direction (CD). One session, 12 neurons. Right, average trajectories from all sessions (\pm s.e.m. bootstrap, Methods). All unperturbed trials (correct and incorrect), grouped by instructed movement. Dotted grey line, decision boundary. Averaging window, 400 ms. **c**, Top, illustration of the CD mode. Middle, activity in ipsilateral and bilateral photoinhibition trials projected onto the CD. All perturbed trials (correct and incorrect), grouped by instructed movement. Dashed blue and red lines, means for unperturbed trials (from **b**). Bottom, behavioural performance in lick-right (blue) and lick-left (red) trials as a function of trajectory distance from the decision boundary. Performance was computed by binning along the CD distance. s.e.m. was obtained by bootstrapping the trials in each bin. **d**, Same as in **c** for activity along persistent mode, which maximizes the difference between perturbed and unperturbed activity at the time of movement onset. This mode does not carry movement-specific information (middle; note that red and blue dashed lines are near each other) and does not predict movement direction (bottom). **e**, Same as in **c** for population activity along the ramping mode, which explains most of the remaining activity variance (Methods). This mode shows robust ramping but is non-selective (middle) and does not predict movement direction (bottom).

We next recorded from left ALM during photoinhibition of left ALM (ipsilateral, one laser spot), right ALM (contralateral, one laser spot), and both hemispheres (four laser spots on each side) ($n = 7$ mice). As before (Fig. 1), spike rate ($>90\%$) (Fig. 2c, d) and selectivity ($>80\%$) (Fig. 2e) recovered 600 ms after ipsilateral perturbation. ALM activity was hardly affected by contralateral photoinhibition (Fig. 2c, d and Extended Data Fig. 2e).

After bilateral photoinhibition, neurons recovered their spike rate on average (Fig. 2c, d), but selectivity failed to recover (Fig. 2c–e). Bilateral photoinhibition with one laser spot produced a larger persistent change in selectivity than unilateral photoinhibition with four laser spots (Extended Data Fig. 7). Recovery of selectivity after

unilateral photoinhibition was less robust with larger photoinhibition size (Extended Data Fig. 7e), similar to behaviour (Fig. 2b). Robustness to perturbation results from redundancy within the bilateral ALM network.

Robustness along the coding direction

Robust systems maintain critical functions in response to perturbations, whereas non-critical features may remain uncorrected²⁶. We analysed population dynamics in the activity space, in which each dimension corresponds to activity of one neuron (6–20 neurons recorded simultaneously; average, 11 neurons; 16 sessions)³⁸. Preparatory activity for different movements (lick-left versus lick-right) corresponded to distinct trajectories in the activity space (Fig. 3a).

We decomposed activity into several modes. First, we estimated the coding direction (CD) along which preparatory activity maximally discriminated upcoming directional licking (Methods, Fig. 3b). After ipsilateral photoinhibition the CD mode recovered to trajectories similar to the unperturbed trials (Fig. 3c; receiver operating characteristic (ROC) values between trajectories at the end of delay epoch: control, 0.76 ± 0.03 ; ipsilateral, 0.73 ± 0.02 ; mean \pm s.e.m. across sessions; Methods). Contralateral photoinhibition had little effect (Extended Data Fig. 8; ROC, 0.74 ± 0.03). As expected (Fig. 2), trajectories were permanently altered after bilateral photoinhibition, resulting in small separation between the trajectories for different trial types at the time of movement onset (ROC, 0.58 ± 0.03). We used a decision boundary, on the CD that separated the lick-left versus lick-right trials (Methods), to predict upcoming movement on a trial-by-trial basis. Deviations towards the lick-right trajectory predicted more frequent lick-right responses and vice versa (Fig. 3, Extended Data Fig. 8). Activity along the CD predicts trial type.

Second, we obtained a mode that maximized sustained effects of ipsilateral perturbations (persistent mode, Fig. 3d). By construction, the persistent mode was altered by the perturbation, up to and beyond movement onset. However, this projection did not discriminate trial type nor predict behaviour on control trials.

Third, a mode that maximally captured the remaining activity variance, showed non-selective ramping during the delay epoch, did not predict behaviour, and was resistant to unilateral and bilateral perturbations (Fig. 3e; see Extended Data Fig. 8 for a full decomposition of ALM dynamics). This ramping mode could reflect non-specific ‘urgency’³⁹ driven by a source external to ALM.

Preparatory activity is therefore maintained by ALM populations along specific trajectories in a sub-space of neural activity space. Circuit dynamics are actively restored along behaviour-related directions in the activity space, but not along certain non-informative directions^{33,40}.

We next examined ALM population activity after bilateral perturbation and its relationship to behaviour. Individual trials with a deviation towards the lick-right trajectory along the CD predicted more frequent lick-right responses and vice versa (Fig. 4, Extended Data Figs 6b and 9). This analysis shows that even after average selectivity is destroyed by perturbations, ALM population dynamics still dictate upcoming movements.

Contralateral input is required for recovery

Preparatory activity is coupled across the two ALM hemispheres (Figs 2 and 3). The small effect seen on activity with contralateral inhibition (Extended Data Figs 2e and 8d) suggests further that ALM hemispheres function as modules, maintaining preparatory activity independently²⁸. After unilateral perturbation, information from the unperturbed side helps to recover the function of the perturbed side. To test directly the role of contralateral ALM input as the corrective signal, we bisected the ALM corpus callosum ($n = 7$ mice) (Fig. 5a, Methods), sparing pyramidal tract and corticothalamic axons (Extended Data Fig. 10).

Notably, behavioural performance was unaffected (Fig. 5b, control trials, before versus after callosotomy, $P > 0.05$, two-tailed t -test), with normal performance 17 h after callosotomy (Extended Data Fig. 10b).

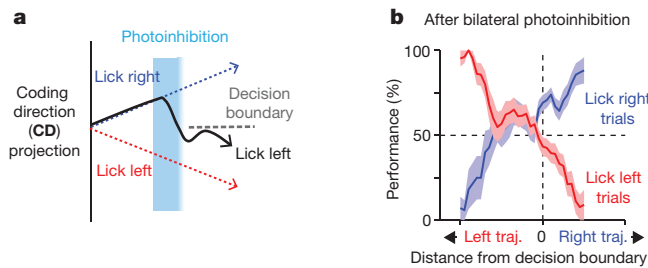


Figure 4 | ALM predicts upcoming movements after bilateral perturbations. **a**, Schematic, using preparatory activity projected onto the coding direction (CD) to predict upcoming movement. **b**, Behavioural performance as a function of trajectory distance from the decision boundary. Same as Fig. 3c for bilateral photoinhibition trials. See Extended Data Fig. 9 for unilateral photoinhibition trials.

However, behavioural performance was now highly sensitive to transient unilateral photoinhibition (Fig. 2b, 1 laser spot: $P = 0.0019$, two-tailed t -test against control). There was a significant interaction between callosotomy and unilateral photoinhibition ($P = 0.0035$, repeated measure two-way analysis of variance, ANOVA). Behavioural performance after unilateral photoinhibition dropped to the same level as bilateral photoinhibition in control mice (Fig. 5b, blue cross).

Preparatory activity in callosotomized mice ($n = 7$ mice) was similar to control mice (Fig. 5c and Extended Data Fig. 10), providing additional evidence that the two ALM hemispheres can maintain preparatory activity independently. After ipsilateral photoinhibition, ALM neurons recovered their average spike rate (Fig. 5c, d), but selectivity failed to recover (Fig. 5e). Selectivity in the coding direction was reduced (Fig. 5f, 16 sessions). Preparatory activity is distributed redundantly across interacting modules in the two ALM hemispheres.

Robust model networks

We compared ALM population dynamics under perturbations (Figs 1 and 2) to predictions from network models (Extended Data Fig. 1). After ipsilateral photoinhibition, ALM activity rapidly recovered to the unperturbed trajectory (Figs 1 and 3). This is inconsistent with attractors with a pair of fixed points (one for each choice condition)²⁹. After release from perturbation, these models decay to the final fixed point and do not return to the trajectory (Extended Data Fig. 4c). Integrator models with a continuum of fixed points generate ramping activity by integrating their inputs^{5,28,37}; these models predict an activity offset compared to the unperturbed trajectories, inconsistent with the data (Extended Data Fig. 1f, g). We also tested randomly connected recurrent networks trained to produce ramping triggered by a transient input^{22–24} (trained random recurrent networks, RRNs). These models failed to recover from perturbations (Extended Data Fig. 1h, i). Overall, all monolithic models consisting of one network were unable to explain robustness.

Preparatory activity is distributed across modules in both ALM hemispheres (Figs 2, 3 and 5). We therefore explored models with the following organizational principles (Fig. 6a): each module can produce ramping independently; recovery from unilateral perturbation is achieved by specific inter-module connectivity (for example, commissural axons); the inter-module connections have little net effect during normal operation. Figure 6b shows a model comprised of two identical modules (corresponding to hemispheres), each consisting of a pair of identical units that inhibit each other and excite themselves to produce ramping activity towards one of two fixed points (representing lick-right or lick-left movements; Methods). Selective commissural connections restored activity on the other side after unilateral transient silencing (Fig. 6b). When the two sides are silenced, the network drifts to one of the fixed points randomly. Similar schemes allowed the integrator and trained RRN to be adapted into a modular and redundant architecture that is robust to unilateral perturbations (Fig. 6c, d; Methods). Imposing modular architecture upon any monolithic model

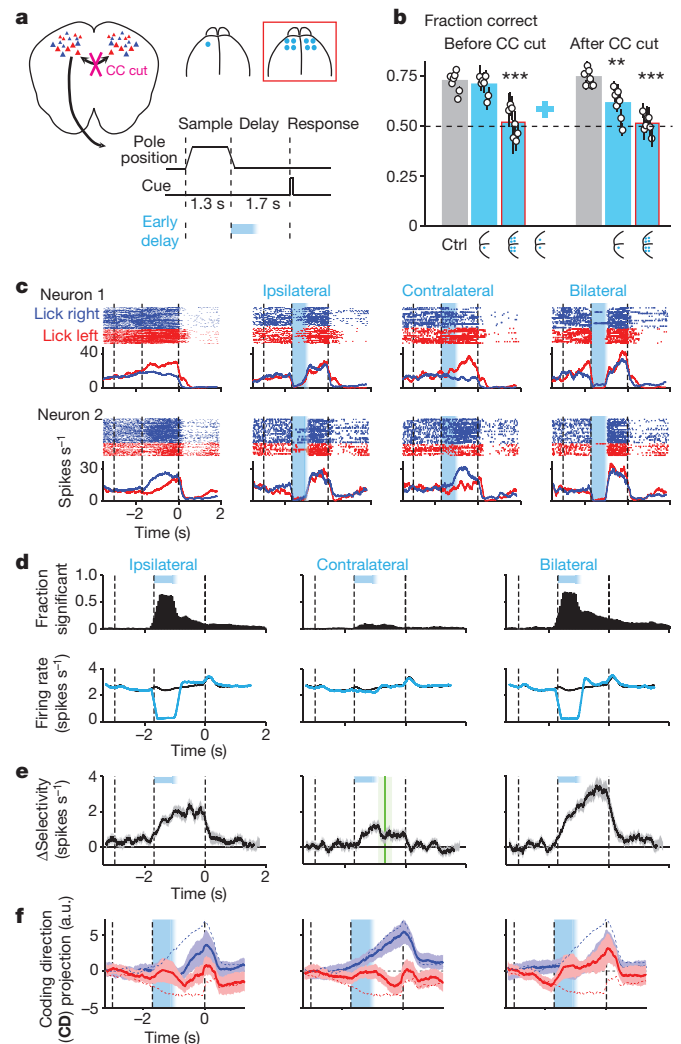


Figure 5 | Contralateral ALM input is required for recovery of preparatory activity. **a**, Left, corpus callosum (CC) bisection. Right, unilateral and bilateral photoinhibition during early delay epochs. **b**, Behavioural performance. Bars, mean. Symbols, individual mice (mean \pm s.e.m., bootstrap). ** $P < 0.01$, *** $P < 0.001$, two-tailed t -test against control. Cyan cross, performance for bilateral photoinhibition, one spot, in a separate group of control mice (data from Fig. 2b). **c**, Two example ALM neurons, after callosotomy. Top, raster plots for 'Lick right' and 'Lick left' trials. Bottom, average spike rate across the population. **d**, Fraction of neurons with significant spike rate change ($n = 325, 322$ and 313). Bottom, average spike rate across the population. **e**, Average change in population selectivity from control ($n = 129$). Same as Fig. 2e. Selectivity recovery: ipsilateral, no recovery; contralateral, 217 ± 228 ms; bilateral, no recovery. **f**, Population activity in photoinhibition trials projected onto the coding direction (CD). Same as Fig. 3c for ipsilateral, contralateral, and bilateral photoinhibition.

allowed it to reproduce the stability found in the data, suggesting the modular architecture itself, and not any particular detail of the models, as the key factor in robustness.

Discussion

Our neurophysiological and behavioural analysis of preparatory activity provides three insights. First, preparatory activity is robust to large, transient perturbations of the network (Fig. 1). Second, unperturbed parts of the network remain functional during the perturbation and help the perturbed part of the network to recover after the perturbation (Figs 2 and 5). Third, premotor cortex preparatory activity recovers in dimensions relevant to behaviour and less so in other dimensions (Fig. 3). This indicates that premotor networks are organized into functionally segregated modules that interact selectively depending on their mutual state (Fig. 6).

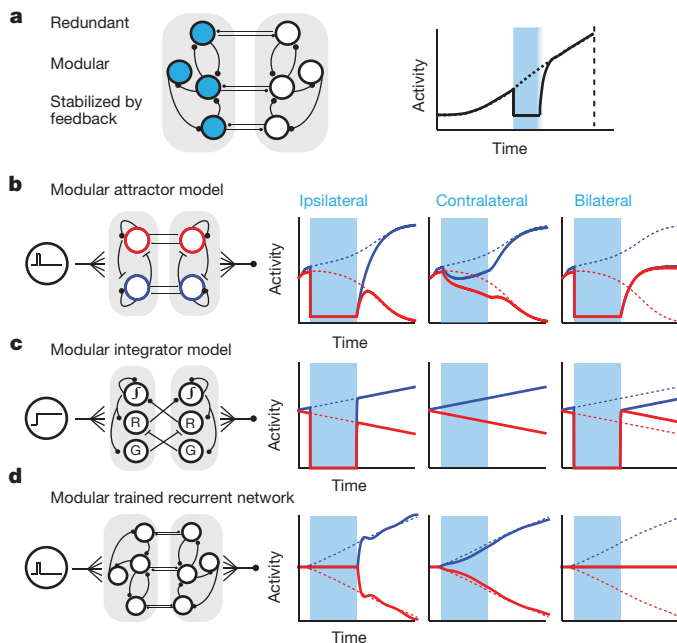


Figure 6 | Modular network models of premotor dynamics. **a**, Schematic, a modular network robust to transient photoinhibition of one module. **b**, Modular attractor model. Neurons with right (blue) and left (red) preferences provide self-excitation and mutual inhibition. Connections between modules involve neurons with similar preference. See Methods for model parameters in **b–d**. **c**, Modular integrator model¹⁵. Connections between modules restore activity on other side ('recovery' neurons, R). Gating (G) neurons cancel the inter-module coupling during normal operation. **d**, Modular recurrent network trained with FORCE learning²² to recapitulate single hemisphere perturbation.

ALM is involved in both planning and driving movements³⁰. Consistent with this view, unilateral photoinhibition of ALM late in the delay epoch abolished the contralateral motor command, resulting in ipsilateral bias (Fig. 1 and Extended Data Fig. 3); furthermore, bilateral photoinhibition during the early delay epoch abolished the motor plan and scrambled future movements (Fig. 2). Previous optogenetic inactivation studies focusing on related brain areas have interpreted a lack of effect of transient inactivation as a lack of role in behaviour^{13,42}. Our results suggest that redundancy across a distributed network could mask possible causal roles in optogenetics experiments.

Modular architecture and functional redundancy are key components of robust engineered systems²⁶. Similarly, our data and previous experiments^{3,30} imply that the cortical networks maintaining motor plans are organized in a redundant and modular fashion. When ALM is silenced in one hemisphere, preparatory activity in the other hemisphere is weakly affected (Fig. 2 and Extended Data Figs 2e and 8d). However, after the perturbation, activity in the unperturbed hemisphere is critical to restore the perturbed preparatory activity in the opposite hemisphere. Preparatory activity is thus distributed in a redundant fashion across functional modules that can both operate independently and correct each other. The cortical networks involved in working memory could be organized in a similar manner^{7–10}. It is likely that modularity and redundancy operate in circuits contained in one hemisphere, perhaps even spatially interdigitated.

The responses of ALM neurons to perturbations can be decomposed into three types of dynamics: modes that are rapidly restored after unilateral perturbation; modes that remain perturbed; and modes that are restored both for unilateral and bilateral perturbation and are thus likely to be driven externally. Only behaviour-relevant modes recovered quickly. ALM responses to perturbations resemble robust systems, in which critical state variables are particularly stiff²⁶. Selective stability of neural dynamics supports the idea that behaviour-related activity

comprises only a low-dimensional subspace of neural activity space^{33,40}, constrained by the structure of neural circuits⁴³. Our findings place constraints on the circuit architectures that underlie memory-related cortical activity and suggest general principles of robust system control in the brain.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 13 October 2015; accepted 8 March 2016.

Published online 13 April 2016.

1. Tanji, J. & Evarts, E. V. Anticipatory activity of motor cortex neurons in relation to direction of an intended movement. *J. Neurophysiol.* **39**, 1062–1068 (1976).
2. Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Ryu, S. I. & Shenoy, K. V. Cortical preparatory activity: representation of movement or first cog in a dynamical machine? *Neuron* **68**, 387–400 (2010).
3. Guo, Z. V. *et al.* Flow of cortical activity underlying a tactile decision in mice. *Neuron* **81**, 179–194 (2014).
4. Erlich, J. C., Bialek, M. & Brody, C. D. A cortical substrate for memory-guided orienting in the rat. *Neuron* **72**, 330–343 (2011).
5. Murakami, M., Vicente, M. I., Costa, G. M. & Mainen, Z. F. Neural antecedents of self-initiated actions in secondary motor cortex. *Nature Neurosci.* **17**, 1574–1582 (2014).
6. Maimon, G. & Assad, J. A. A cognitive signal for the proactive timing of action in macaque LIP. *Nature Neurosci.* **9**, 948–955 (2006).
7. Fuster, J. M. & Alexander, G. E. Neuron activity related to short-term memory. *Science* **173**, 652–654 (1971).
8. Funahashi, S., Bruce, C. J. & Goldman-Rakic, P. S. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J. Neurophysiol.* **61**, 331–349 (1989).
9. Romo, R., Brody, C. D., Hernandez, A. & Lemus, L. Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature* **399**, 470–473 (1999).
10. Liu, D. *et al.* Medial prefrontal activity during delay period contributes to learning of a working memory task. *Science* **346**, 458–463 (2014).
11. Wang, X. J. Decision making in recurrent neuronal circuits. *Neuron* **60**, 215–234 (2008).
12. Gold, J. I. & Shadlen, M. N. The neural basis of decision making. *Annu. Rev. Neurosci.* **30**, 535–574 (2007).
13. Hanks, T. D. *et al.* Distinct relationships of parietal and prefrontal cortices to evidence accumulation. *Nature* **520**, 220–223 (2015).
14. Mainen, Z. F. & Sejnowski, T. J. Reliability of spike timing in neocortical neurons. *Science* **268**, 1503–1506 (1995).
15. Cannon, S. C., Robinson, D. A. & Shamma, S. A proposed neural network for the integrator of the oculomotor system. *Biol. Cybern.* **49**, 127–136 (1983).
16. Sheffield, M. E., Best, T. K., Mensh, B. D., Kath, W. L. & Spruston, N. Slow integration leads to persistent action potential firing in distal axons of coupled interneurons. *Nature Neurosci.* **14**, 200–207 (2011).
17. Yoshida, M. & Hasselmo, M. E. Persistent firing supported by an intrinsic cellular mechanism in a component of the head direction system. *J. Neurosci.* **29**, 4945–4952 (2009).
18. Barak, O., Sussillo, D., Romo, R., Tsodyks, M. & Abbott, L. F. From fixed points to chaos: three models of delayed discrimination. *Prog. Neurobiol.* **103**, 214–222 (2013).
19. Murakami, M. & Mainen, Z. F. Preparing and selecting actions with neural populations: toward cortical circuit mechanisms. *Curr. Opin. Neurobiol.* **33**, 40–46 (2015).
20. Fisher, D., Olasagasti, I., Tank, D. W., Aksay, E. R. & Goldman, M. S. A modeling framework for deriving the structural and functional architecture of a short-term memory microcircuit. *Neuron* **79**, 987–1000 (2013).
21. Wang, X. J. Synaptic reverberation underlying mnemonic persistent activity. *Trends Neurosci.* **24**, 455–463 (2001).
22. Sussillo, D. & Abbott, L. F. Generating coherent patterns of activity from chaotic neural networks. *Neuron* **63**, 544–557 (2009).
23. Laje, R. & Buonomano, D. V. Robust timing and motor patterns by taming chaos in recurrent neural networks. *Nature Neurosci.* **16**, 925–933 (2013).
24. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
25. London, M., Roth, A., Beeren, L., Hausser, M. & Latham, P. E. Sensitivity to perturbations *in vivo* implies high noise and suggests rate coding in cortex. *Nature* **466**, 123–127 (2010).
26. Kitano, H. Biological robustness. *Nature Rev. Genet.* **5**, 826–837 (2004).
27. Csete, M. E. & Doyle, J. C. Reverse engineering of biological complexity. *Science* **295**, 1664–1669 (2002).
28. Aksay, E. *et al.* Functional dissection of circuitry in a neural integrator. *Nature Neurosci.* **10**, 494–504 (2007).
29. Kopec, C. D., Erlich, J. C., Brunton, B. W., Deisseroth, K. & Brody, C. D. Cortical and subcortical contributions to short-term memory for orienting movements. *Neuron* **88**, 367–377 (2015).
30. Li, N., Chen, T. W., Guo, Z. V., Gerfen, C. R. & Svoboda, K. A motor cortex circuit for motor planning and movement. *Nature* **519**, 51–56 (2015).

31. Komiyama, T. *et al.* Learning-related fine-scale specificity imaged in motor cortex circuits of behaving mice. *Nature* **464**, 1182–1186 (2010).
32. Shenoy, K. V., Sahani, M. & Churchland, M. M. Cortical control of arm movements: a dynamical systems perspective. *Annu. Rev. Neurosci.* **36**, 337–359 (2013).
33. Kaufman, M. T., Churchland, M. M., Ryu, S. I. & Shenoy, K. V. Cortical activity in the null space: permitting preparation without movement. *Nature Neurosci.* **17**, 440–448 (2014).
34. Tanaka, M. Cognitive signals in the primate motor thalamus predict saccade timing. *J. Neurosci.* **27**, 12109–12118 (2007).
35. Seung, H. S. How the brain keeps the eyes still. *Proc. Natl Acad. Sci. USA* **93**, 13339–13344 (1996).
36. Amit, D. J. & Brunel, N. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cereb. Cortex* **7**, 237–252 (1997).
37. Lim, S. & Goldman, M. S. Balanced cortical microcircuitry for maintaining information in working memory. *Nature Neurosci.* **16**, 1306–1314 (2013).
38. Stopfer, M., Jayaraman, V. & Laurent, G. Intensity versus identity coding in an olfactory system. *Neuron* **39**, 991–1004 (2003).
39. Cisek, P., Puskas, G. A. & El-Murr, S. Decisions in changing conditions: the urgency-gating model. *J. Neurosci.* **29**, 11560–11571 (2009).
40. Druckmann, S. & Chklovskii, D. B. Neuronal circuits underlying persistent representations despite time varying activity. *Curr. Biol* **22**, 2095–2103 (2012).
41. Goldman, M. S. Memory without feedback in a neural network. *Neuron* **61**, 621–634 (2009).
42. Diester, I. *et al.* An optogenetic toolbox designed for primates. *Nature Neurosci.* **14**, 387–397 (2011).
43. Sadtler, P. T. *et al.* Neural constraints on learning. *Nature* **512**, 423–426 (2014).

Acknowledgements We thank B. DePasquale, A. Finkelstein, D. Gutnisky, A. Hantman, H. Inagaki, V. Jayaraman, J. Magee, S. Peron, S. Romani and N. Spruston for comments on the manuscript and discussion, T. Pluntke for animal training, A. Hu for histology, T. Harris and B. Barbarits for silicon probe recording system. This work was funded by Howard Hughes Medical Institute. N.L. and K.D. are Helen Hay Whitney Foundation postdoctoral fellows.

Author Contributions N.L., K.S. and S.D. conceived and designed the experiments. N.L. and K.D. performed behavioural experiments. N.L. performed electrophysiology and optogenetic experiments. K.D. and S.D. performed modeling. N.L., K.D., K.S. and S.D. analysed data and wrote the paper.

Author Information Data have been deposited at the CRCNS (<https://crcns.org/>) and can be accessed at <http://dx.doi.org/10.6080/KORB72JW>. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to K.S. (svobodak@janelia.hhmi.org) or S.D. (druckmanns@janelia.hhmi.org).

METHODS

Mice. This study is based on data from 33 mice (both males and females, age >postnatal day (P) 60). Ten VGAT-ChR2-EYFP mice (Jackson Laboratory, JAX 014548) and nine PV-ires-cre⁴⁴ crossed to Rosa26-LSL-ReaChR, red-shifted channelrhodopsin reporter mice (JAX 24846)⁴⁵, were used for photoinhibition behaviour experiments. A subset of these mice (five VGAT-ChR2-EYFP mice, seven PV × ReaChR mice) was used for simultaneous electrophysiology and behaviour. Seven mice (six VGAT-ChR2-EYFP, one PV × ReaChR mice) were used for the callosotomy experiment. Two Tlx_{PL56}-cre (MMRRC 036547)⁴⁶ crossed to Ai32 (Rosa26-ChR2 reporter mice, JAX 012569)⁴⁷ mice were used for photoactivation experiment. Two untrained VGAT-ChR2-EYFP mice and two untrained PV × ReaChR mice were used to characterize the photoinhibition in ALM. One Tlx_{PL56}-cre mouse was used for anatomical characterization of the ALM axonal projection pattern.

All procedures were in accordance with protocols approved by the Janelia Institutional Animal Care and Use Committee. Mice were housed in a 12h:12h reverse light:dark cycle and tested during the dark phase. On days not tested, mice received 1 ml of water. On other days, mice were tested in experimental sessions lasting 1–2 h, in which they received all their water (range, 0.5–2 ml). If mice did not maintain a stable body weight, they received supplementary water⁴⁸. All surgical procedures were carried out aseptically under 1–2% isoflurane anaesthesia. Buprenorphine HCl (0.1 mg kg⁻¹, intraperitoneal injection; Bedford Laboratories) was used for postoperative analgesia. Ketoprofen (5 mg kg⁻¹, subcutaneous injection; Fort Dodge Animal Health) was used at the time of surgery and postoperatively to reduce inflammation. After the surgery, mice were allowed to recover for at least 3 days with free access to water before water restriction.

The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Behaviour and surgery. Mice were prepared for photoinhibition and electrophysiology with a clear-skull cap and a headpost³. The scalp and periosteum over the dorsal surface of the skull were removed. A layer of cyanoacrylate adhesive (Krazy glue, Elmer's Products Inc.) was directly applied to the intact skull. A custom made headpost⁴⁸ was placed on the skull with its anterior edge aligned with the suture lambda (approximately over cerebellum) and cemented in place with clear dental acrylic (Lang Dental Jet Repair Acrylic; 1223-clear). A thin layer of clear dental acrylic was applied over the cyanoacrylate adhesive covering the entire exposed skull, followed by a thin layer of clear nail polish (Electron Microscopy Sciences, 72180).

The behavioural task and training have been described^{3,48}. The stimulus was a metal pin (0.9 mm in diameter), presented at one of two possible positions (Fig. 1a). The two pole positions were 4.29 mm apart along the anterior-posterior axis (approximately 40° of whisking angle) and were constant across sessions. The posterior pole position was 5 mm from the whisker pad. A two-spout lickport (4.5 mm between spouts) was used to record answer licks and deliver water rewards.

At the beginning of each trial, the vertical pole moved into reach of the whiskers (0.2 s travel time), where it remained for 1 s, after which it was retracted (retraction time 0.2 s). The sample epoch is defined as the time between the pole movement onset to 0.1 s after the pole retraction onset (sample epoch, 1.3 s, Fig. 1a). Mice touched the object at both pole positions, typically with a different set of whiskers. The delay epoch (durations, 1.2–1.7 s) followed the sample epoch. An auditory 'go' cue indicated the end of the delay epoch (pure tone, 3.4 kHz, 0.1 s duration). Licking early during the trial was punished by a loud alarm sound (siren buzzer, 0.05 s duration), followed by a brief timeout (1–1.2 s). Licking the correct lickport after the go cue led to a liquid reward (3 µl). Licking the incorrect lickport triggered a timeout (2–5 s). Trials in which mice did not lick within a 1.5-s window after the go cue were rare and typically occurred at the end of a session.

Photoinhibition. Light from a 473 nm laser (Laser Quantum, Gem 473) or a 594 nm laser (Cobolt Inc., Cobolt Mambo 100) was controlled by an acousto-optical modulator (AOM; Quanta Tech) and a shutter (Vincent Associates). Photoinhibition of ALM was performed through the clear-skull cap implant by directing the laser over the skull (beam diameter: 400 µm at 4σ). The light transmission through the intact skull is 50%³. Photoinhibition was deployed on 25% of the behavioural trials during behavioural testing. To prevent the mice from distinguishing photoinhibition trials from control trials using visual cues, a 'masking flash' (40 1-ms pulses at 10 Hz) was delivered using 470 nm or 591 nm LEDs (Luxeon Star) near the eyes of the mice. The masking flash began as the pole started to move and continued through the end of photoinhibition. For silencing we stimulated cortical GABAergic neurons in VGAT-ChR2-EYFP mice, or parvalbumin-positive interneurons in PV-ires-cre mice crossed to reporter mice expressing ReaChR⁴⁵. The two methods resulted in similar photoinhibition (Extended Data Fig. 2). The photoinhibition silenced 90% of spikes (Extended Data Fig. 2b) in a cortical area of 1 mm radius (at half-max) through all cortical layers³ (Extended

Data Fig. 2d). To minimize rebound excitation after photoinhibition offset, we linearly ramped down the laser power (100 or 200 ms). This photostimulus was empirically determined³ to produce robust photoinhibition with minimal rebound (Extended Data Fig. 2c).

The duration of the delay epoch varied to accommodate different photoinhibition conditions. In the unilateral photoinhibition experiment (Fig. 1, Extended Data Fig. 2a), a fixed 1.3-s delay epoch was used. We used a 40-Hz photostimulus with a sinusoidal temporal profile (1.5 mW average power) and a 100-ms linear ramp. We photoinhibited for 0.5 s, including the 100 ms ramp, during different task epochs (Fig. 1a). Photostimuli ended 1.6 s (sample), 0.8 s (early delay), 0.3 s (late delay), or 0 s (before cue) before the go cue. We also tested unilateral photoinhibition of longer durations in separate experiments (1.3 s including a 100-ms ramp, Extended Data Fig. 3). To accommodate the longer photoinhibition, we randomly varied the duration of the delay epoch from 1.2 s to 1.7 s in 0.1-s increments. This resulted in photoinhibition that terminated at different times before the go cue.

In the bilateral photoinhibition experiment (Fig. 2), a fixed 1.7-s delay epoch was used to allow more time for neuronal activity to recovery after photoinhibition. We photoinhibited for 0.8 s, including a 200-ms ramp during offset, either at the beginning of the sample epoch or at the beginning of the delay epoch. To photoinhibit single cortical locations (Fig. 2a, 1 laser spot), we used the 40-Hz sinusoidal photostimulus (1.5 mW average power). To photoinhibit multiple cortical locations (Fig. 2a, multiple laser spots), we used a constant photostimulus and a scanning galvo (GVSM002, Thorlabs), which stepped the laser beam sequentially through the photoinhibition sites at the rate of 1 step per 5 ms (step time: <4.8 ms; measured using a photodiode). Peak power was adjusted depending on the number of cortical locations to achieve 1.5 mW average power per location. The photoinhibition during scanning was similar to the standard condition (Extended Data Fig. 2).

To estimate the proportion of ALM silenced by photoinhibition, we estimated the boundaries of ALM using photoinhibition behavioural data from³. ALM was defined as the area where photoinhibition over the entire delay epoch produced significant behavioural effects. ALM boundaries (Fig. 1b, grey area) were derived by deconvolving the area producing significant behavioural effects with the point-spread function of the photoinhibition method³ (Extended Data Fig. 2d). At 80% activity reduction, photoinhibition with 1 laser spot covered 58% of ALM in one hemisphere (Fig. 1b).

Photoactivation. For photoactivation we stimulated layer 5 intratelencephalic neurons in Tlx_{PL56}-cre mice⁴⁶ crossed to reporter mice expressing ChR2 (Ai32)³⁰. The delay epoch was 1.3 s long. The photostimulus was a 20-Hz sinusoid (0.53 mW average power) applied during different task epochs (Extended Data Fig. 5b). Photoactivation was deployed on 40% of the behavioural trials during electrophysiology.

Electrophysiology. A small craniotomy (diameter, 1 mm) was made over left ALM (centred on 2.5 mm anterior, 1.5 mm lateral) one day before the recording session³. Extracellular spikes were recorded using NeuroNexus silicon probes (A4x8-5 mm-100-200-177). The 32 channel voltage signals were multiplexed, digitized by a PCI6133 board at 312.5 kHz (National instrument) at 14 bit, demultiplexed (sampling at 19531.25 Hz) and stored for offline analysis. Three to seven recordings were made from each craniotomy. To minimize brain movement, a drop of silicone gel (3-4680, Dow Corning) was applied over the craniotomy after the electrode was in the tissue. The tissue was allowed to settle for several minutes before the recording started.

During electrophysiology, photoinhibition was deployed on 40% of the trials to obtain a larger number of trials per condition. Three photoinhibition conditions were tested during each recording session. In the unilateral photoinhibition experiment (Fig. 1, Extended Data Fig. 2a), photoinhibition during sample, early delay, and late delay epoch were tested. In the bilateral photoinhibition experiment (Fig. 2, Extended Data Fig. 2a), photoinhibition of left ALM (ipsilateral, 1 laser spot), right ALM (contralateral, 1 laser spot), and both hemispheres (4 laser spot) were tested. In separate experiments (Extended Data Figs 2a and 7), ipsilateral photoinhibition with 4 laser spots, contralateral photoinhibition with 4 laser spots, and bilateral photoinhibition with 1 laser spot were tested.

Callosotomy. The placement of the corpus callosum cut was determined based on ALM axonal projection patterns. AAV2/1-CAG-EGFP (Addgene, plasmid 28014) was injected into one hemisphere of ALM (Extended Data Fig. 10c). The injection coordinate was 2.5 mm anterior to bregma and 1.5 mm lateral to the midline. The injection was made through the thinned skull using a custom volumetric injection system. Glass pipettes (Drummond) were pulled and bevelled to a sharp tip (outer diameter of 30 µm). Pipettes were back-filled with mineral oil and front-loaded with viral suspension immediately before injection. 50-nl volumes were injected 500 and 800 µm deep. Two weeks after injection, mice were perfused and their brains were sectioned (50 µm) and processed using standard

fluorescent immunohistochemical techniques. Confocal images were acquired on a Zeiss microscope, a 10 × objective and a Hamamatsu Orca Flash 4 camera⁴⁶.

ALM axons extend caudally from the injection site. Corpus callosum axons separate from pyramidal tract and corticothalamic axons approximately 1.2 mm anterior to bregma. ALM corpus callosum axons were confined to the anterior regions of corpus callosum and were densest around 1 mm from bregma (Extended Data Fig. 10c). Corpus callosum axon bisection was made through an elongated craniotomy either over the left (3 mice) or right (4 mice) hemisphere. A 3.5-mm-deep cut was made using a micro knife (Fine Science Tools, 10318-14) mounted on a micromanipulator (Sutter Instrument). The cut was 0.5 mm from the midline and was at a slight angle to avoid the pyramidal tract and corticothalamic axons (Extended Data Fig. 10d). The cut extended from 1.5 mm anterior to bregma to 1 mm posterior. Care was taken to avoid damaging the superior sagittal sinus. In the same surgery, a second craniotomy was made over left ALM for electrophysiology. Approximately 17 h after the surgery mice were tested in behavioural experiments (Fig. 5, Extended Data Fig. 10). Mice were tested in daily recording sessions for 5–7 days after the callosotomy. Mice were perfused immediately after the last recording session and the brains were processed for histology (Extended Data Fig. 10d). In a subset of the mice, brain sections were stained for GFAP (mouse; Sigma G3893, 1:2,000 dilution) (Extended Data Fig. 10d).

Behavioural data analysis. Performance was computed as the fraction of correct reports, excluding lick-early trials (Figs 1–5). Chance performance was 50%. We also separately computed the performance for lick-right and lick-left trials (Figs 3, 4 and Extended Data Figs 3, 6, 9). Behavioural effects of photoinhibition were quantified by comparing the performance under photoinhibition with control performance using two-tailed *t*-test (Figs 1, 2, 5 and Extended Data Fig. 3).

Electrophysiology data analysis. The extracellular recording traces were band-pass filtered (300–6 kHz). Events that exceeded an amplitude threshold (4 s.d. of the background) were subjected to manual spike sorting to extract single units³. 1,012 single units were recorded during behaviour across 58 recording sessions (20 sessions of unilateral experiments, Fig. 1; 38 sessions of bilateral experiments, Fig. 2, Extended Data Fig. 7). Spike widths were computed as the trough-to-peak interval in the mean spike waveform. Units with spike width <0.35 ms were defined as fast-spiking neurons (72 out of 1,012) and units with spike widths >0.45 ms as putative pyramidal neurons (890 out of 1,012). Units with intermediate values (0.35–0.45 ms, 50 out of 1,012) were excluded. This classification was previously verified by optogenetic tagging of GABAergic neurons³. We concentrated our analyses on the putative pyramidal neurons.

Neurons were tested for significant trial-type selectivity during the sample, delay, or response epochs, using the spike counts from the lick-left and lick-right trials (two-tailed *t*-test, $P < 0.05$). Neurons that significantly differentiated trial types during any one of the trial epochs were deemed 'selective' (634 out of 890). To compute selectivity (Figs 1, 2, 5 and Extended Data Fig. 1), we first determined each neuron's preferred trial type using spike counts from a subset of the trials (10 trials), selectivity is calculated as the spike rate difference between the trial types on the remaining data. Standard errors of the mean were obtained by bootstrap across neurons.

To quantify the effect of photoinhibition on individual ALM neuron spike rates (Figs 1, 2, 5 and Extended Data Figs 5, 7), we used a two-tailed *t*-test on spike counts binned in 400-ms windows (control versus photoinhibition). Spike counts from lick-right trials and lick-left trials were pooled. Spike rates were tested at different times during the task (in 50-ms time steps) and significance was reported for $P < 0.01$.

Quantification of the effects of perturbations on movement selectivity was complicated by the fact that ALM selectivity is coupled to upcoming movements. Grouping trials by the final movement (for example, using only correct lick-right trials) to compute selectivity would miss the trials in which photoinhibition caused the mice to switch future movements, thus underestimating the effects of photoinhibition on selectivity. We therefore used all trials (correct and incorrect) to compute selectivity when quantifying selectivity changes caused by photoinhibition (Figs 1, 2, 5 and Extended Data Figs 5, 7). Selectivity change was the selectivity difference between control and photoinhibition trials. To quantify the recovery time course of selectivity after photoinhibition, we looked for the first time bin when selectivity on photoinhibition trials reached 80% of the control selectivity (Figs 1g, 2e, 5e and Extended Data Figs 5, 7, green lines). Standard errors of the mean were obtained by bootstrap across neurons.

Analysis of population dynamics in the activity space. To analyse the relationship between ALM population activity and upcoming movements, we restricted analysis to the recording sessions from the bilateral photoinhibition experiments (Fig. 2) with >5 neurons recorded simultaneously for >5 trials per condition (16 out of 38 sessions, Figs 3, 4 and Extended Data Figs 6, 8, 9). For a population of n neurons, we found an $n \times 1$ vector, in the n dimensional activity space that maximally separated the response vectors in lick-right trials and lick-left trials, we term this vector the coding direction (CD).

Average spike counts were computed in a 400-ms window in 10-ms steps. For each movement direction (lick right and lick left, correct trials only) we computed the average spike counts $\bar{\mathbf{x}}_{\text{lick right}}$ and $\bar{\mathbf{x}}_{\text{lick left}}$, $n \times 1$ response vectors that described the population response at that time. During the sample and delay epochs the direction of the difference in the mean response vectors, $\mathbf{w}_t = \bar{\mathbf{x}}_{\text{lick right}} - \bar{\mathbf{x}}_{\text{lick left}}$, was stable (correlation of \mathbf{w}_t values between late sample epoch versus late delay epoch, 0.61 ± 0.05 ; Extended Data Fig. 9b). We averaged the \mathbf{w}_t values from the sample and delay epochs to obtain the coding direction (CD). Because our estimate of the covariance was noisy, the CD gave better discrimination than the linear discriminant vector (CD divided by the within-group covariance).

The projection along the CD captured $65.6 \pm 5.1\%$ of the population selectivity for lick-left and lick-right trials over the sample and delay epochs (root mean square (r.m.s.), of the spike rate difference between lick-right trials and lick-left trials), and $36.4 \pm 6.3\%$ of the total variance in ALM task-related activity (Extended Data Fig. 8a). Activity variance was quantified as the r.m.s. of the baseline subtracted activity over the sample and delay epoch.

To project the ALM population activity along the CD we used independent control and perturbation trials from the trials used to compute the CD. For each trial we computed the spike counts for each neuron, \mathbf{x} ($n \times 1$), at each time point. The projected trajectories in Figs 3, 5 and Extended Data Figs 6–9 were obtained as $\mathbf{CD}^T \mathbf{x}$. Both correct and incorrect trials were used to compute the projected trajectories, grouped by the instructed movements. To quantify the separation between trajectories on lick-right and lick-left trials, we computed ROC values using $\mathbf{CD}^T \mathbf{x}$ at the end of the delay epoch for each session. To average trajectories across multiple behavioural sessions (Figs 3, 5 and Extended Data Figs 7–9), we first offset the trajectories for a particularly session by subtracting the mean $\mathbf{CD}^T \mathbf{x}$ across all trials and time points in that session. This removed fluctuations in mean activity from session to session. The offsets were computed using the independent control trials that were used to calculate the CD. Standard errors of the mean were obtained by bootstrapping individual sessions.

To predict upcoming movements using ALM responses projected onto the CD (Figs 3, 4 and Extended Data Figs 8b, 9), we used the response vector \mathbf{x} from the last time bin before the go cue (last 400 ms of the delay epoch). For each session, we computed a decision boundary (DB) to best separate the projected responses, $\mathbf{CD}^T \mathbf{x}$, from lick-right and lick-left trials:

$$\text{DB} = \frac{\mathbf{CD}^T \mathbf{x}_{\text{lick right}} / \sigma_{\text{lick right}}^2 + \mathbf{CD}^T \mathbf{x}_{\text{lick left}} / \sigma_{\text{lick left}}^2}{1 / \sigma_{\text{lick right}}^2 + 1 / \sigma_{\text{lick left}}^2}$$

σ^2 is the variance of the projected responses $\mathbf{CD}^T \mathbf{x}$ across multiple lick-right or lick-left trials. Both the CD and decision boundary were computed using independent control trials and separate control and photoinhibition trials were used to predict performance. Data from multiple sessions were pooled in Figs 3, 4 and Extended Data Fig. 9.

We decomposed ALM activity into three forms of dynamics (Fig. 3 and Extended Data Fig. 8). The modes were computed using a subset of control trials (correct trials only) and ipsilateral perturbation trials. The projections in the figures are for independent control trials and perturbation trials. The projection along the CD (mode 1) captured the movement selectivity in activity. The persistent mode (mode 2) was the difference in the mean response vectors between ipsilateral perturbed and unperturbed lick-right trials at the go cue. Mode 3 was the mean response vectors between ipsilateral perturbed and unperturbed lick-left trials at the go cue, further rotated using Gram–Schmidt process to be orthogonal to mode 2. We did not orthogonalize the CD mode and persistent mode, so that any potential selectivity common to these modes was not removed. There was a small overlap between mode 1 and modes 2–3 (the activity variance and selectivity shared by modes 1–3 are quantified in Extended Data Fig. 8a). Modes 2 and 3 describe the vast majority of the persistent changes in activity after ipsilateral perturbations.

Two additional modes (4 and 5) captured the remaining activity variance. We first found eigenvectors of the population activity matrix using singular value decomposition. The data for the singular value decomposition (SVD) was an $n \times t$ matrix, consisting of the baseline-subtracted PSTHs for n neurons, with the lick-right and lick-left trials concatenated together (t time bins). The first two eigenvectors ($n \times 1$) were rotated using the Gram–Schmidt process to be orthogonal to modes 1–3, yielding modes 4 and 5. Modes 1–5 together explained $98.5 \pm 0.5\%$ of the total variance of task-related activity and $95.8 \pm 1.2\%$ of population selectivity over the sample and delay epochs. To predict upcoming movements using the projected responses on persistent mode and ramping mode (Fig. 3), we computed decision boundaries on the projected responses using the same procedures as for the CD mode.

Modelling and simulation. Model code can be found at <https://github.com/kpdaie/LiDaie>.

We constructed neural networks that have the ability to produce slow ramps of preparatory activity (Fig. 6a) when receiving transient or constant input, similar to a subset of ALM neurons. Our models include a phenomenological attractor model (Fig. 6b), explicit integrators (Fig. 6c and Extended Data Fig. 1f, g), and recurrent neural networks (RNNs) trained to produce ramping output (Fig. 6d and Extended Data Fig. 2h, i). We compared the responses of the models and ALM to transient silencing.

All networks were simulated for two seconds. Photoinhibition was simulated by holding the activity of half of the neurons in each network at zero for times $0.2\text{ s} < t < 1.0\text{ s}$. Activity of the i th neuron $r_i(t)$ was governed by the equation:

$$\tau \frac{dr_i(t)}{dt} = -r_i(t) + \sum_{j=1}^N W_{i,j} f(r_j(t)) + I_i(t) + T_i(t) + \xi_i(t)$$

The cellular time constant, τ , the connectivity matrix, W , and the synaptic nonlinearity, $f(r)$, differed across the models. N is the number of neurons, $T_i(t)$ is a tonic and non-selective input, and $\xi_i(t)$ is Gaussian random noise. In all simulations networks received either transient ($0.05\text{ s} < t < 0.1\text{ s}$) or persistent ($0.1\text{ s} < t < 1.9\text{ s}$) sensory inputs $I_i(t)$.

Simple integrator model (Extended Data Fig. 1f). The network was simulated with $N = 100$, $\tau = 100\text{ ms}$, and linear synapses, $f(r) = r$. The connectivity matrix was constructed so that all eigenvalues except for one were equal to zero. The non-zero eigenvalue was set to 0.99, producing feedback so that the activity of the network decays with time constant $\tau / (1 - 0.99) = 10\text{ s}$ (ref. 35). The input was either persistent (Extended Data Fig. 1f, left) or transient (Extended Data Fig. 1f, right), in which case the output from the integrator was cascaded into a second identical network to produce ramping activity. Silencing was simulated by holding the activity of a randomly-selected population of 50 neurons at zero for times $0.2\text{ s} < t < 1.0\text{ s}$.

Integrator with corrective feedback (Extended Data Fig. 1g). Corrective feedback³⁷ was incorporated into an integrator network to confer robustness against perturbations. The model consists of a pair of excitatory and inhibitory neurons. Corrective feedback was achieved by a mismatch in the time constants for excitatory and inhibitory connections, which generates negative derivative feedback. The network exhibits robustness against random perturbations that equally affect the excitatory and inhibitory neurons, but is not robust against asymmetric activation of inhibitory neurons (for example, photoinhibition). The function $f(r)$ is linear $f(r) = s$ where the auxiliary variable s is determined by the equation:

$$\tau_{\text{syn},i,j} \frac{ds_{i,j}}{dt} = -s_{i,j} + r_j(t)$$

The synaptic time constant $\tau_{\text{syn},i,j}$ determines how quickly the post-synaptic currents respond to changes in presynaptic activity. The synaptic time constants were: inhibitory synapses, 10 ms; excitatory to inhibitory neurons, 25 ms; excitatory to excitatory neurons, 100 ms. The network received a task-selective persistent input. Photoinhibition was simulated by injecting large currents into the inhibitory neuron and disallowing negative spike rates, which results in silencing of the excitatory neuron.

Trained RNN, FORCE learning (Extended Data Fig. 1h). We used FORCE²² training to minimize the difference between the network readout ($z(t)$) and a ramping waveform (Extended Data Fig. 1h). $z(t)$ is a linear combination of the activity of each neuron with weights determined by the vector \mathbf{w}_o (that is, $z(t) = \sum \mathbf{w}_o \cdot \mathbf{r}(t)$). Tuning of $z(t)$ was accomplished by simulating the activity of an initially randomly connected recurrent neural network (RNN) for 2 seconds (time step, 1 ms) and adjusting W every 2 ms during the simulation. This process was repeated 30 times.

The initial connectivity matrix was chosen to be sparse with a connection probability $p = 0.1$. Non-zero connections were chosen from a Gaussian random distribution. The variance in connection strength was $\frac{1.5^2}{pN}$. 1.5 is a gain factor which is sufficiently strong to produce chaotic activity⁴⁹. In addition, we used $\tau = 200\text{ ms}$, $f(r) = \tanh(r)$, $N = 400$ and transient input. Photoinhibition was simulated by transiently clamping the activity of a randomly-selected population of 200 (that is, $N/2$) neurons to zero. The network received either persistent (Extended Data Fig. 1h, left) or transient (Extended Data Fig. 1h, right) sensory input. For persistent input the network behaved similar to an integrator exhibiting a recovery of selectivity, albeit at an offset level upon removal of photoinhibition.

Trained RNN, tamed chaos (Extended Data Fig. 1i). RNNs were trained with FORCE as described above and further stabilized (tamed chaos)²³. The algorithm was designed to stabilize selected trajectories in chaotic networks via a recursive retuning of recurrent connection strengths based on a recursive least-squares rule⁵⁰. To minimize the number of synapses that required tuning, the FORCE network was made sparse by eliminating weak connections that were smaller than

an arbitrary threshold and using linear regression to adjust the remaining weights to maintain the dynamics. Elimination of weak synapses reduced the time needed to train the network. Twenty iterations of the tamed chaos algorithm were then run with weights being adjusted every 10 ms. Perturbations were applied as described for the FORCE trained network above. This training resulted in a modest increase in the robustness of the network.

Modular attractor (Fig. 6b). Two identical two-neuron unilateral attractor modules were constructed so that each neuron excited itself with weight 0.5235 and inhibited the other neuron in the same module with -0.5235 . Each neuron was reciprocally connected with one partner from the other module with strength 0.3. $\tau = 100\text{ ms}$ and $f(r) = g(r) - g(0)$, where $g(r) = \frac{1.4}{1 + e^{-(r-0.5)/0.3}}$. Transient input $I_i(t)$ (amplitude, 0.1) was provided to either the right-preferring (blue, Fig. 6b) or left-preferring (red, Fig. 6b) neurons, depending on the trial type. All neurons received a tonic input $T_i(t)$ with amplitude 0.5 and noise $\xi_i(t)$ with variance 0.01. **Modular integrator (Fig. 6c).** Two modules with inter-module connections were tuned to produce robustness against unilateral photoinhibition. Each module consisted of four neurons (numbers refer to neuronal indices, with reference to the connection matrix W): Right preferring integrator neurons (1, 5) and left preferring integrator neurons (2, 6). Integration was produced by positive feedback achieved through mutual inhibition between left and right preferring neurons with strength -1 (ref. 15); these integrating pairs are represented schematically by the circles labelled \int in Fig. 6c. The modules are connected through the recovery neurons (3, 7; 'R' in Fig. 6c) and gating inhibitory neurons (4, 8; 'G' in Fig. 6c). The input $I_i(t)$ was persistent with amplitude 0.04 to the right-preferring neuron and -0.04 to the left-preferring neuron during lick-right trials. The signs of the inputs were flipped for lick-left trials. In addition, each integrator neuron received tonic input $T_i(t)$ with amplitude 40.0 to produce baseline activity at 20.0.

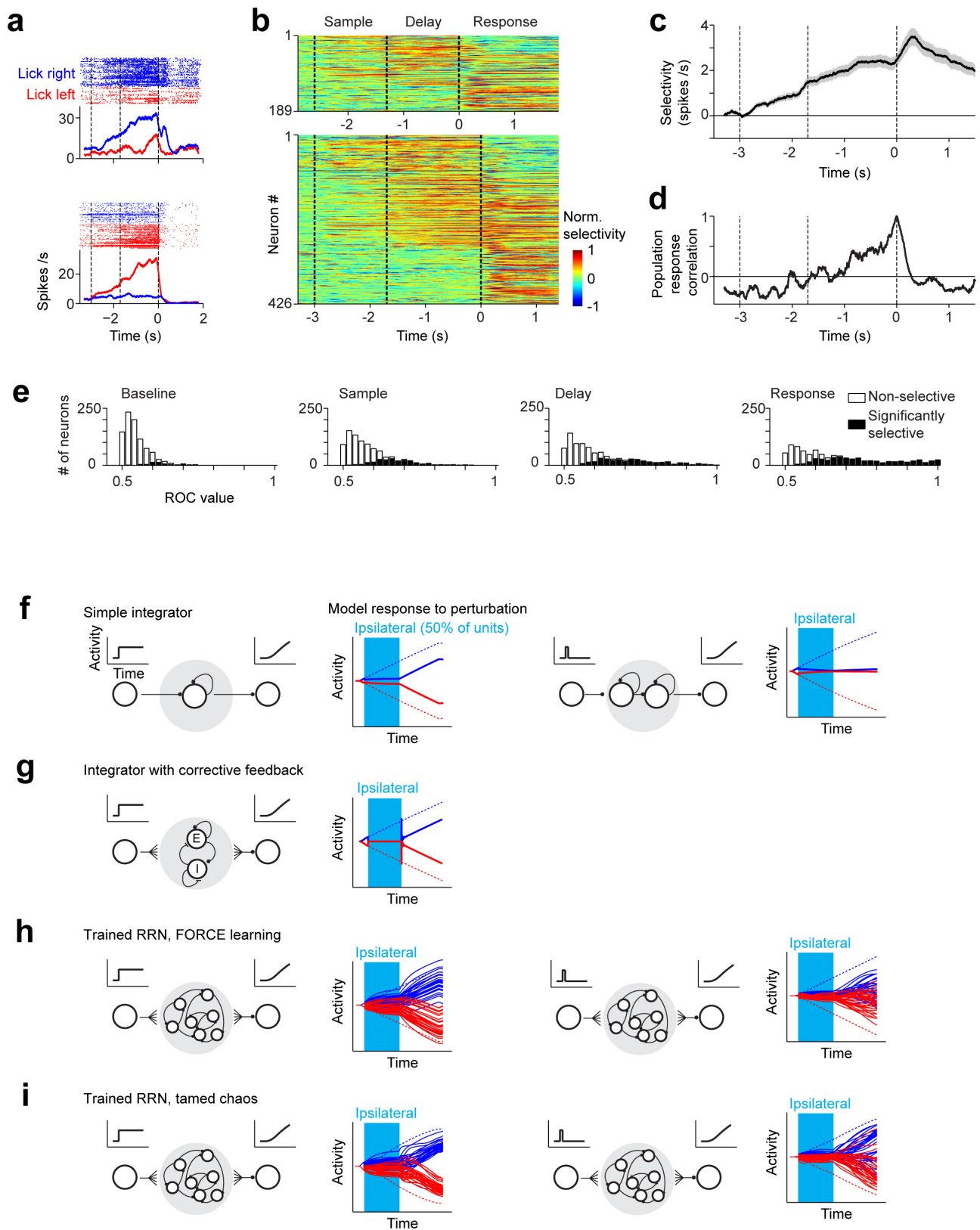
The recovery (Fig. 6c) and gating inhibitory (Fig. 6c) neurons together produce robustness. They receive positive input from the right-preferring neuron and negative input from the left-preferring neuron. After removal of photoinhibition, the recovery neuron restores the activity of the contralateral integrator neurons. This restorative connection has strength 0.5. To avoid excessive coupling between modules during normal function the recovery neuron is strongly inhibited by the gating neuron with strength -6.0 . The full connectivity matrix is shown below. For example, element $W_{1,7}$ is the connection from the recovery neuron in module 2 (neuron 7) onto the right preferring neuron of module 1 (neuron 1).

$$W = \begin{pmatrix} 0 & -1 & 0 & 0 & 0 & 0 & 0.5 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & -0.5 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & -6 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & -0.5 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -6 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \end{pmatrix}$$

The time constant of this network was $\tau = 10\text{ ms}$. Synapses in this network were linear, but activity was restricted to be positive. $\xi_i(t) = 0$ in this network.

Modular tamed chaos (Fig. 6d). To generate a modular RNN we started, as above (see Trained RNN, FORCE learning), with a randomly connected RNN with $N = 400$. We then classified 200 neurons as module 1 and the other 200 as module 2. FORCE training was performed as described above, but we first tuned only the intra-modular connections so that each module could produce its own ramping output. Next, inter-modular connections were trained in the presence of transient photoinhibition (described above) of module 1, so that the output of module 1 would recover upon removal of photoinhibition and the output of module 2 would be minimally affected by the photoinhibition. This process was then repeated for photoinhibition of module 2. In this network $T_i(t) = 0$ and $\xi_i(t) = 0$.

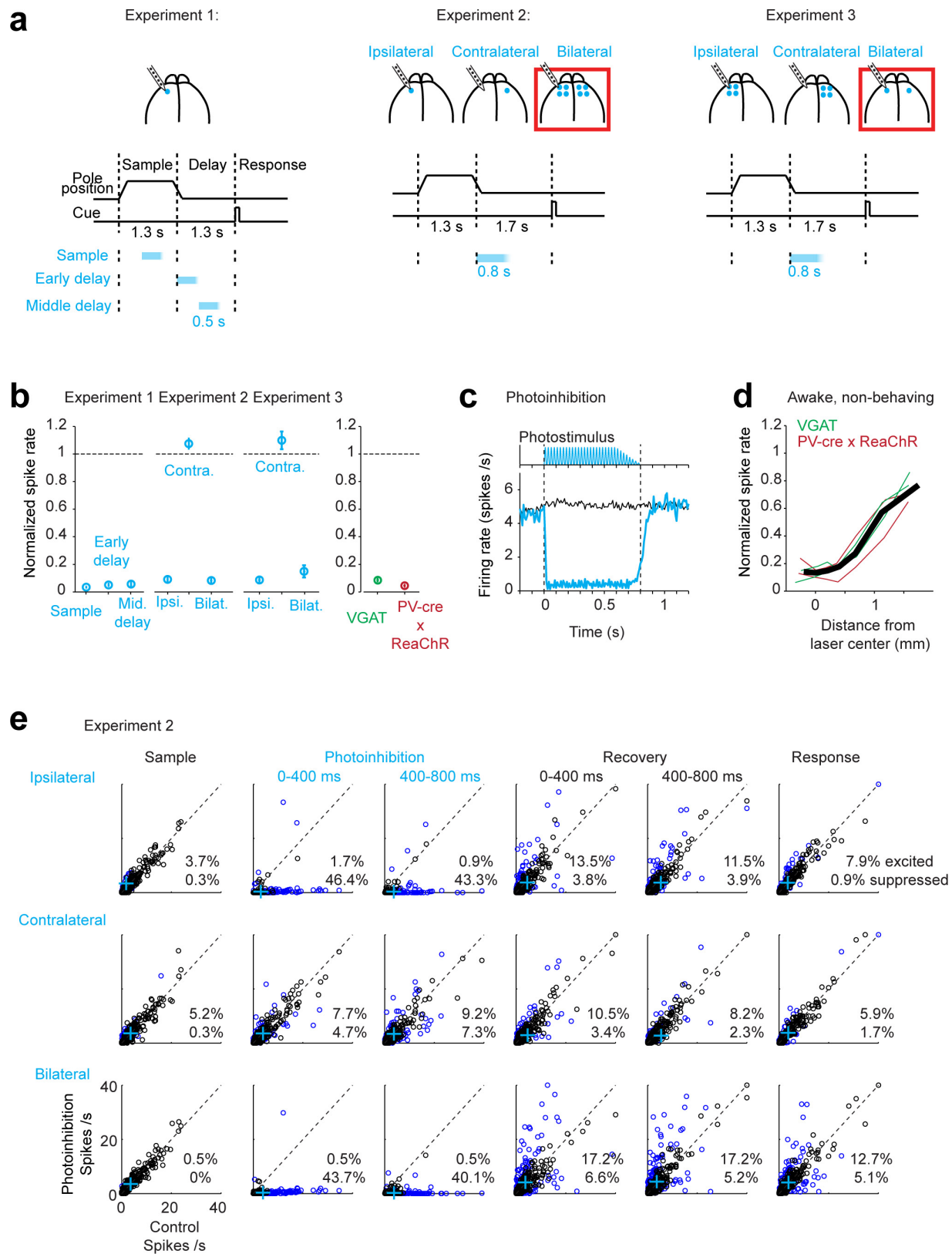
44. Hippenmeyer, S. *et al.* A developmental switch in the response of DRG neurons to ETS transcription factor signaling. *PLoS Biol.* **3**, e159 (2005).
45. Hooks, B. M., Lin, J. Y., Guo, C. & Svoboda, K. Dual-channel circuit mapping reveals sensorimotor convergence in the primary motor cortex. *J. Neurosci.* **35**, 4418–4426 (2015).
46. Gerfen, C. R., Paletzki, R. & Heintz, N. GENSAT BAC cre-recombinase driver lines to study the functional organization of cerebral cortical and basal ganglia circuits. *Neuron* **80**, 1368–1383 (2013).
47. Madisen, L. *et al.* A toolbox of Cre-dependent optogenetic transgenic mice for light-induced activation and silencing. *Nature Neurosci.* **15**, 793–802 (2012).
48. Guo, Z. V. *et al.* Procedures for behavioral experiments in head-fixed mice. *PLoS ONE* **9**, e88678 (2014).
49. Sompolinsky, H., Crisanti, A. & Sommers, H. J. Chaos in random neural networks. *Phys. Rev. Lett.* **61**, 259–262 (1988).
50. Haykin, S. *Adaptive Filter Theory* 4th edn (Prentice Hall, 2002).



Extended Data Figure 1 | See next page for caption.

Extended Data Figure 1 | ALM activity during motor planning and network models of premotor dynamics. **a**, Two example ALM neurons with selectivity during the object location discrimination task, out of 890 putative pyramidal neurons from 12 mice (Methods). Correct lick-right (blue) and lick-left (red) trials only. Dashed lines demarcate behavioural epochs. Averaging window, 200 ms. **b**, ALM population selectivity. Top, delay epoch was 1.3 s; bottom, delay epoch was 1.7 s. Selectivity is the difference in spike rate between the preferred and non-preferred trial type, normalized to the peak selectivity (Methods). Only putative pyramidal neurons with significant trial selectivity are shown ($n = 634$ out of 890). In addition, neurons tested for <15 trials for each trial type (19 out of 634) were excluded. **c**, Average population selectivity in spike rate (black line, \pm s.e.m. across neurons, bootstrap). **d**, Population response correlation. Pearson's correlation between the population response vectors at different times during the task and the population response vector at

the onset of the go cue (time = 0). All selective putative pyramidal neurons were used, even if not recorded at the same time (ignoring potential correlations between neurons). To equalize the contributions of individual neurons, each neuron's response was mean-subtracted and normalized to the variance of its response across the entire trial (computed in time bins of 200 ms). **e**, Distribution of selectivity across the population during different epochs. For each neuron, a ROC value between lick-right and lick-left trials was computed using the spike counts during the particular behavioural epoch. Solid bars, neurons with significant trial-type selectivity ($P < 0.05$, two-tailed t -test using spike counts). **(f-i)** Monolithic models (see Methods). Each solid line represents the activity of the network's output in response to photoinhibition. Activity does not recover after transiently silencing subsets of neurons in: Simple integrator model³⁵ (**f**), Integrator with corrective feedback³⁷ (**g**), Trained RNN, FORCE learning²² (**h**), Trained RNN, Tamed Chaos²³ (**i**).

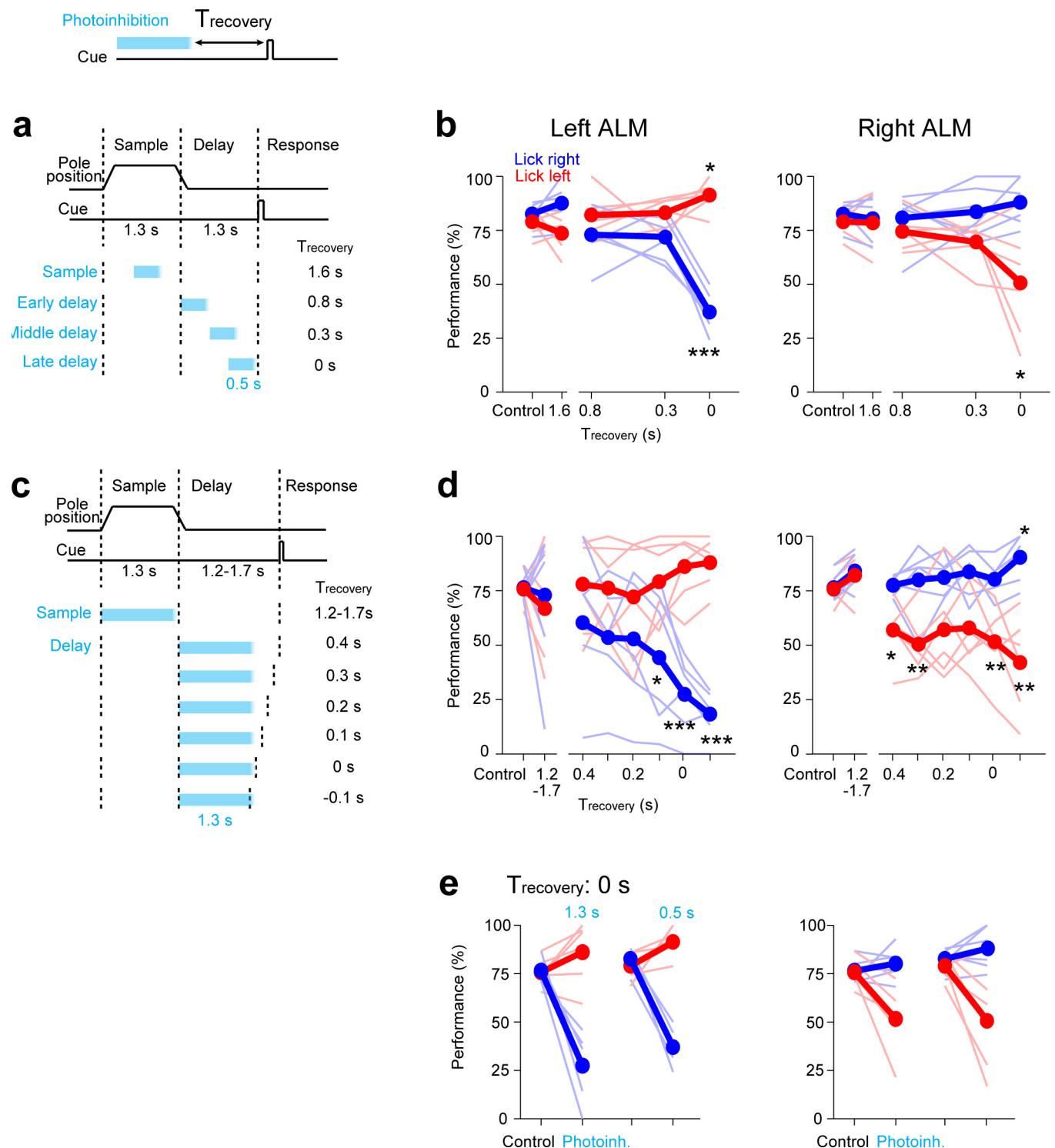


Extended Data Figure 2 | See next page for caption.

Extended Data Figure 2 | Characterization of photoinhibition.

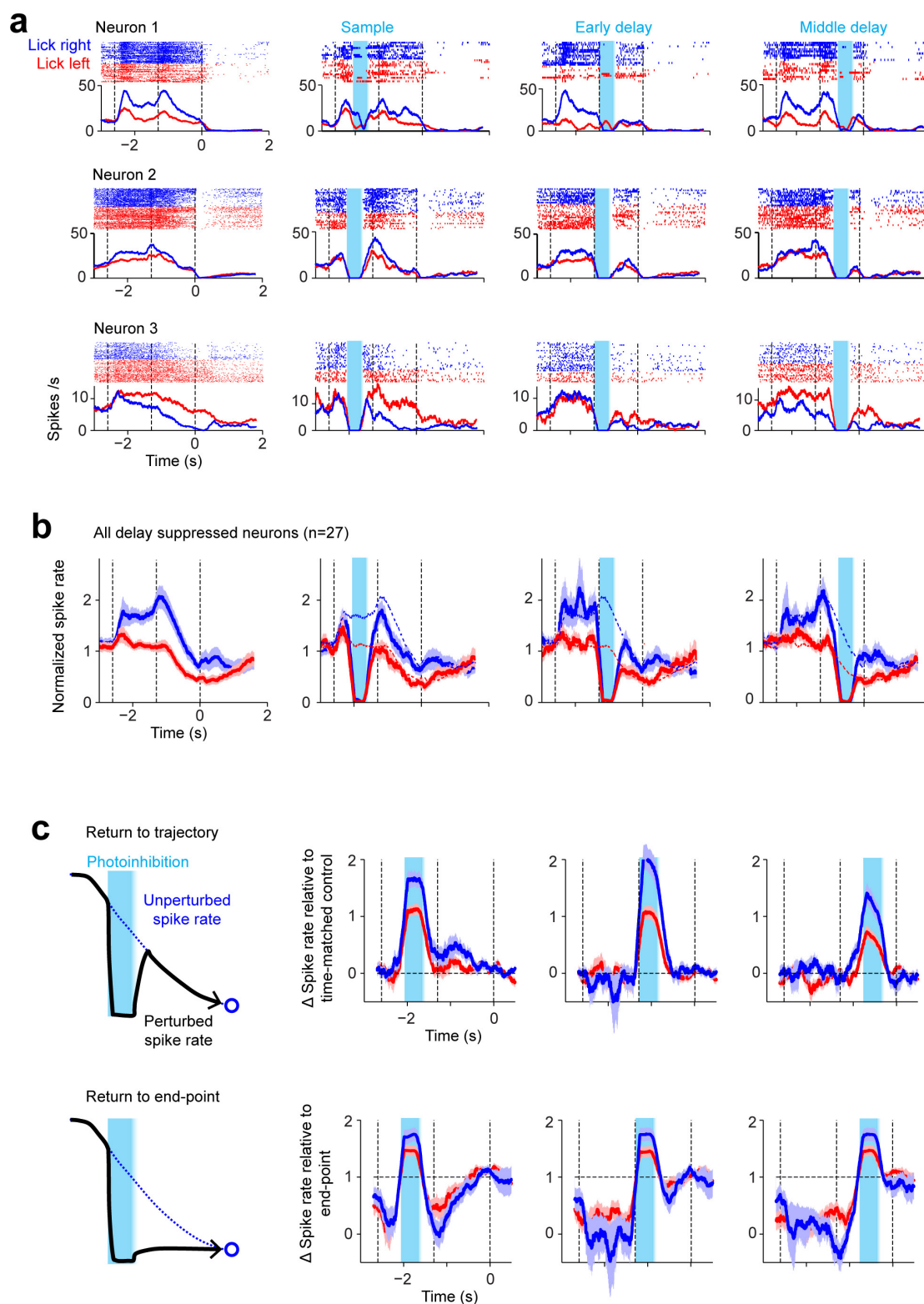
a, Silicon probe recording and photoinhibition in different experimental configurations used in this study. Experiment 1, data presented in Fig. 1 and Extended Data Figs 3, 4; experiment 2, data presented in Figs 2–4 and Extended Data Figs 6, 8, 9; experiment 3, data presented in Extended Data Fig. 7. **b**, Effect of photoinhibition on putative pyramidal neurons. For each neuron, spike rate during photoinhibition was normalized to spike rate in control trials. Left, experiment 1: $n = 117$, 110 and 109 neurons from 6 mice; experiment 2: $n = 300$, 294 and 301 from 7 mice; experiment 3: $n = 52$, 52 and 102 from 3 mice. Ipsilateral and bilateral photoinhibition similarly silenced neuronal activity. Average spike rate across the population was little affected by contralateral photoinhibition. Right, comparison of photoinhibition in VGAT-ChR2-EYFP mice and PV-ires-cre mice crossed to a ReaChR reporter line (Methods)⁴⁵. Photoinhibition was similar in the two mouse lines ($>90\%$ activity reduction). Data from ipsilateral photoinhibition from experiment 2 ($n = 94$ neurons from 3 VGAT mice; $n = 201$ from 4 PV-cre \times ReaChR mice). Error bars, s.e.m. over neurons. Neurons with mean spike rate of <1 spikes s^{-1} were excluded. **c**, Top, photostimuli were shaped to minimize rebound activity after photoinhibition. Peak photostimulus intensity was gradually reduced over 200 ms during stimulus offset. Bottom, average spike rate across the population (black, control; cyan, photoinhibition). Data from experiment 2,

ipsilateral photoinhibition, $n = 300$ neurons from 7 mice. **d**, Effect of photoinhibition versus distance from the laser centre under the standard photostimulus (1 laser spot). Neurons were pooled across cortical depths. Recording data were obtained from ALM of 4 untrained mice under awake and non-behaving conditions. Recording procedures were described previously³. Thin lines, individual mice ($n = 246$ neurons, 2 VGAT-ChR2-EYFP mice, 2 PV-ires-cre \times ReaChR mice). **e**, Average spike rates on control versus photoinhibition lick-right trials during different epochs of the task. Data from experiment 2. Photoinhibition was for 800 ms at the beginning of the delay epoch. The delay epoch was 1.7 s. Columns from left to right: the last 400 ms of the sample epoch, the first 400 ms of the photoinhibition, the last 400 ms of the photoinhibition, the first 400 ms after photoinhibition, 400–800 ms after photoinhibition, first 400 ms of the response epoch (see **a** for trial structure). Top, ipsilateral photoinhibition (1 laser spot, Methods); middle, contralateral photoinhibition (1 laser spot); bottom, bilateral photoinhibition (4 laser spots). Coloured dots, neurons with significant spike rate change ($P < 0.01$, two tailed t -test). Crosses, population means. No rebound excitation was detected after photoinhibition offset on average (**d**). A small proportion of neurons showed rebound excitation which was balanced by a low level of sustained inhibition in a larger proportion of neurons. Results are similar for lick-left trials (not shown).



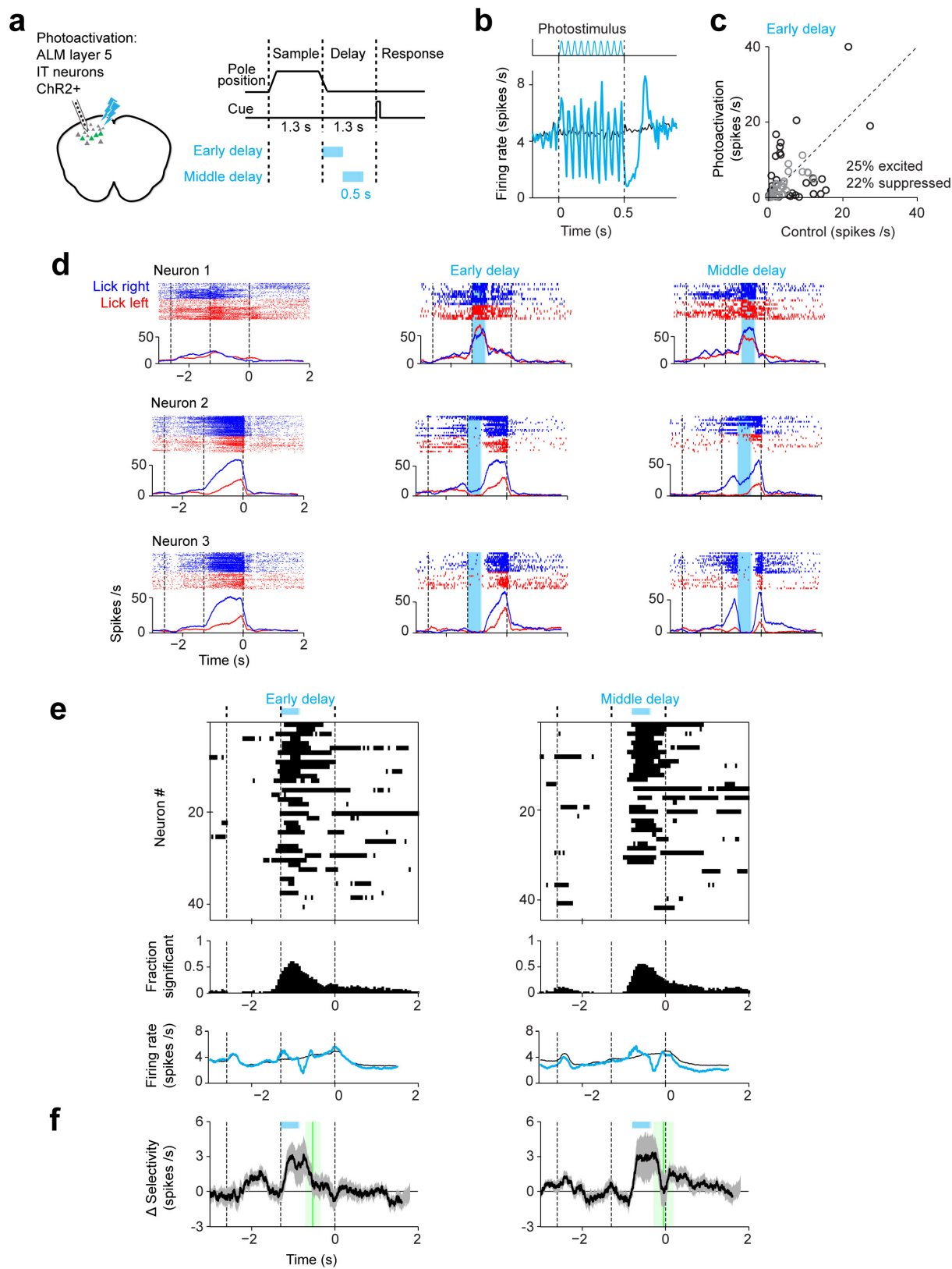
Extended Data Figure 3 | Unilateral photoinhibition of ALM immediately before movement causes ipsilateral bias. **a**, Unilateral photoinhibition of ALM during different task epochs. Sample epoch, 1.3 s; delay epoch, 1.3 s. Photoinhibition, 0.5 s (0.4 s and 0.1 s ramp, Methods). **b**, Performance with 0.5 s photoinhibition of left or right ALM during different trial epochs. Performance was plotted as a function of time interval between photoinhibition offset (the end of ramp offset) and the onset of go cue (T_{recovery}). Performance was not significantly affected for $T_{\text{recovery}} > 0.3$ s. Thick lines, mean; thin lines, individual mice ($n = 5$). * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, two-tailed t -test.

c, Unilateral photoinhibition of ALM during different task epochs. Sample epoch, 1.3 s; delay epoch, variable duration, 1.2–1.7 s in 0.1-s increments. Trials with different delay epoch durations were randomly interleaved. Photoinhibition was for 1.3 s (1.2 s and 0.1 s ramp, Methods), resulting in different T_{recovery} . **d**, Performance with 1.3 s photoinhibition. Plot is similar to **b**. Performance was not significantly affected for $T_{\text{recovery}} > 0.3$ s. **e**, Photoinhibition (0.5 s) immediately before the go cue is similar to the behavioural effect caused by photoinhibition during the entire delay epoch (1.3 s). Photoinhibition data at $T_{\text{recovery}} = 0$ from **b** and **d** was re-plotted.



Extended Data Figure 4 | ALM neurons with decreasing spike rates during the delay epoch recovered their normal spike rates after unilateral photoinhibition. a, Three example ALM neurons with decreasing spike rates during the delay epoch. Top, spike raster. Bottom, PSTH. All lick-right (blue) and lick-left (red) trials. Dashed lines, behavioural epochs. Blue shades, photoinhibition. **b**, Normalized spike rate for all neurons with significant spike rate decrease at the end of the delay epoch compared to the beginning of the delay epoch ($P < 0.05$, two-tailed t -test; 400 ms windows; pooled across trial types). 27 neurons from 6 mice. The spike rate for each neuron was normalized to the mean spike rate. Blue, preferred trial type; red, non-preferred. Mean \pm s.e.m. across

neurons, bootstrap. Dotted lines, spike rates in control trials. **c**, The data are consistent with a return to the normal trajectory and inconsistent with decay to the end point. Top, spike rate difference between perturbed trials and the time-matched spike rates in control trials. Bottom, spike rate difference between perturbed trials and the spike rates at the end of the delay epoch in control trials. Data from **b**. Mean \pm s.e.m. across neurons, bootstrap. Spike rate difference relative to time-matched control show significantly smaller root mean squared error (r.m.s.e.) than spike rate difference relative to end point ($P < 0.001$, paired t -test). r.m.s. was computed during the epoch between photoinhibition offset and the go cue.

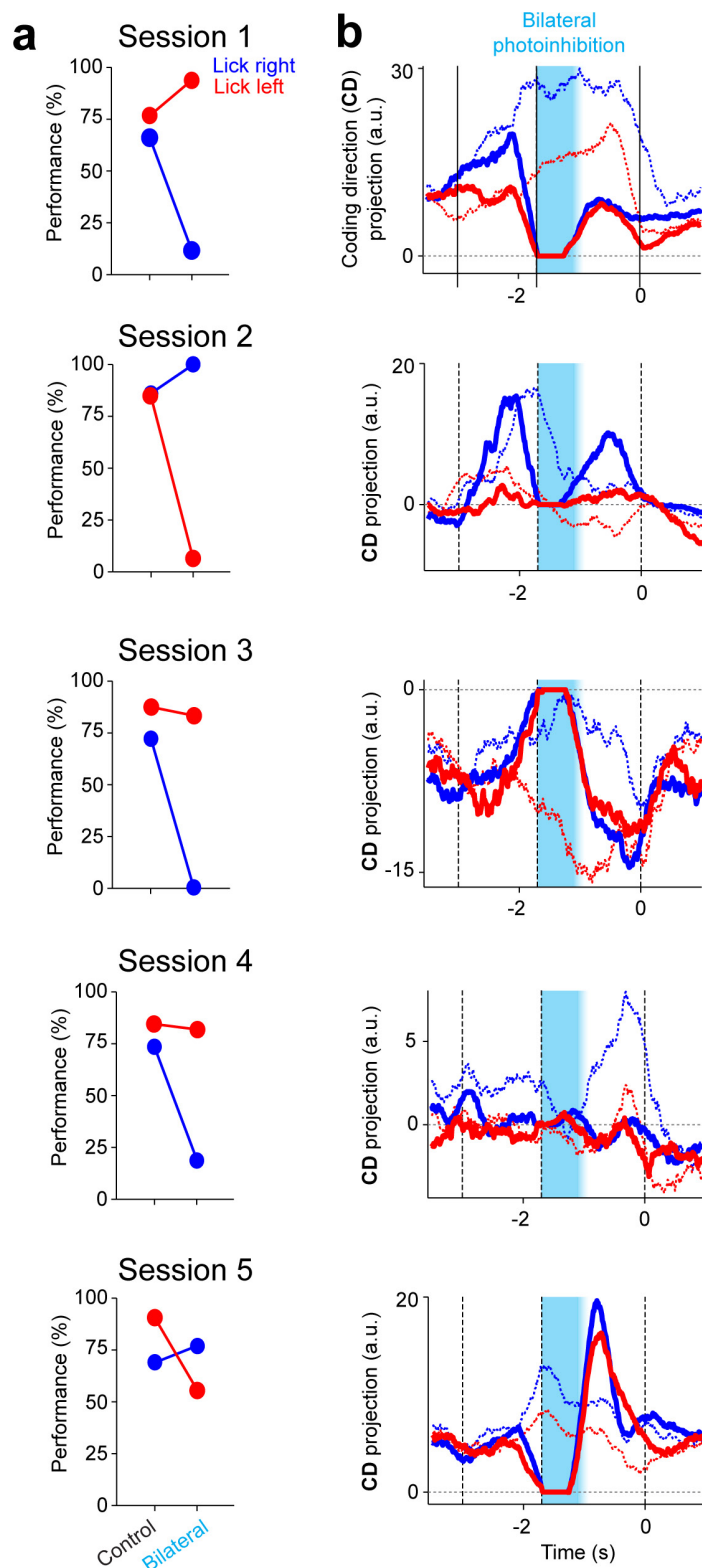


Extended Data Figure 5 | See next page for caption.

Extended Data Figure 5 | Preparatory activity is robust to

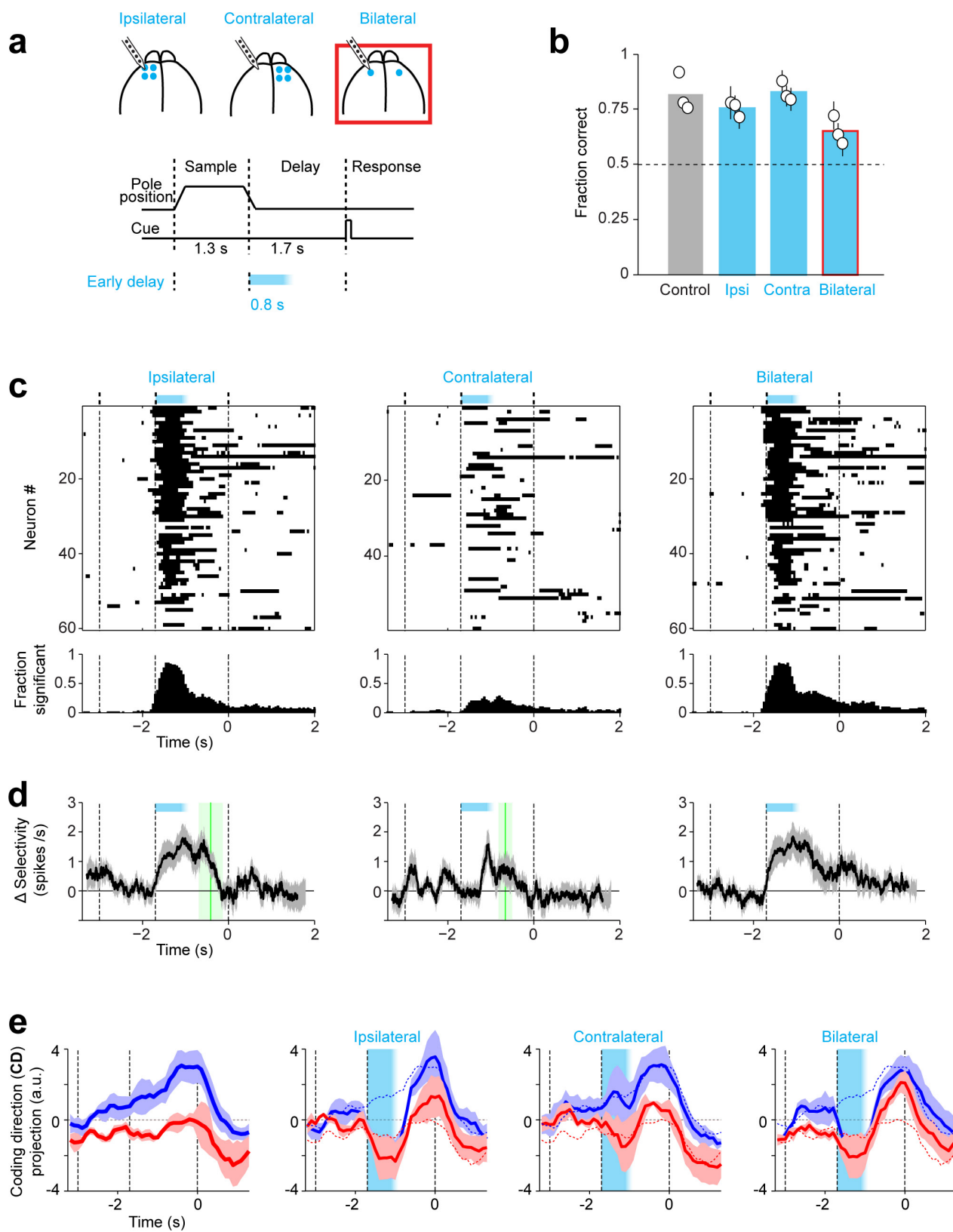
photoactivation. **a**, Left, silicon probe recording during unilateral photoactivation of a subset of excitatory neurons. *Tlx₁PL56-Cre* mice were crossed to Ai32 (*Rosa26-ChR2*) reporter mice to express ChR2 in layer 5 intratelencephalic (IT) neurons⁵⁰. Right, task structure and timing of photoactivation (cyan). **b**, Top, photostimulus. Bottom, average spike rate across the population ($n = 69$ neurons from 2 mice). Black, control; cyan, photoactivation. Rebound inhibition was observed after photoactivation. **c**, Effect of photoactivation on spike rates. Data are for photoactivation during early delay epoch. Black circles, neurons with significant spike rate change ($P < 0.01$, two tailed t -test). Photoactivation during sample epoch: 19% excited, 22% suppressed; late delay epoch: 15% excited, 17% suppressed. Lick-right and lick-left trials were pooled to compute spike rates. **d**, Three example ALM neurons. Top, spike raster. Bottom, PSTH.

All lick-right (blue) and lick-left (red) trials. Dashed lines, behavioural epochs. Blue shades, photoinhibition. **e**, Top, significant spike rate changes relative to control are highlighted for individual neurons. Neurons (rows) are sorted based on their mean spike rate across the trial epochs. Neurons with mean spike rate below 1 spikes s^{-1} or tested for less than 3 trials are excluded. Middle, fraction of neurons with significant spike rate change ($n = 43, 44$ from 2 mice). Bottom, average spike rate across the population. **f**, Average population selectivity change from control (Δ selectivity \pm s.e.m. across neurons, bootstrap). Only selective neurons tested for >3 trials in all conditions are shown ($n = 26$). Green lines, time points when the selectivity recovered to 80% of control selectivity (mean \pm s.e.m. across neurons, bootstrap). Sample epoch: 249 ± 68 ms to recover to 80% of control selectivity; early delay: 275 ± 168 ms; middle delay: 250 ± 218 ms.



Extended Data Figure 6 | ALM dynamics predicts upcoming movements at the level of behavioural sessions. **a**, Behavioural performance on control and bilateral photoinhibition trials. **b**, Time course of activity trajectories projected onto the coding direction (CD). Dotted lines, average trajectories from control lick-right (blue) and lick-left (red) trials. Solid lines, average trajectories from bilateral photoinhibition trials. Each plot shows data from one session for one mouse. Trajectories in photoinhibition trials were similar to control trials before photoinhibition and were persistently altered by transient bilateral photoinhibition. The resultant trajectories were inconsistent from session

to session: in some cases the altered trajectories were closer to the lick-right control trajectories (blue dotted lines), and in other cases closer to the lick-left control trajectories (red dotted lines). Averaging window, 400 ms. In sessions with altered activity trajectories that were closer to the control lick-left trajectories, movements were biased to the left, resulting in high performance in lick-left trials and low performance in lick-right trials (session 1, 4). The opposite behavioural bias was observed when altered activity trajectories were closer to the control lick-right trajectories (session 2, 3, 5). The biases in movement were predicted based ALM activity trajectories. Session 1–5, $n = 20, 16, 18, 10$ and 12 neurons.



Extended Data Figure 7 | See next page for caption.

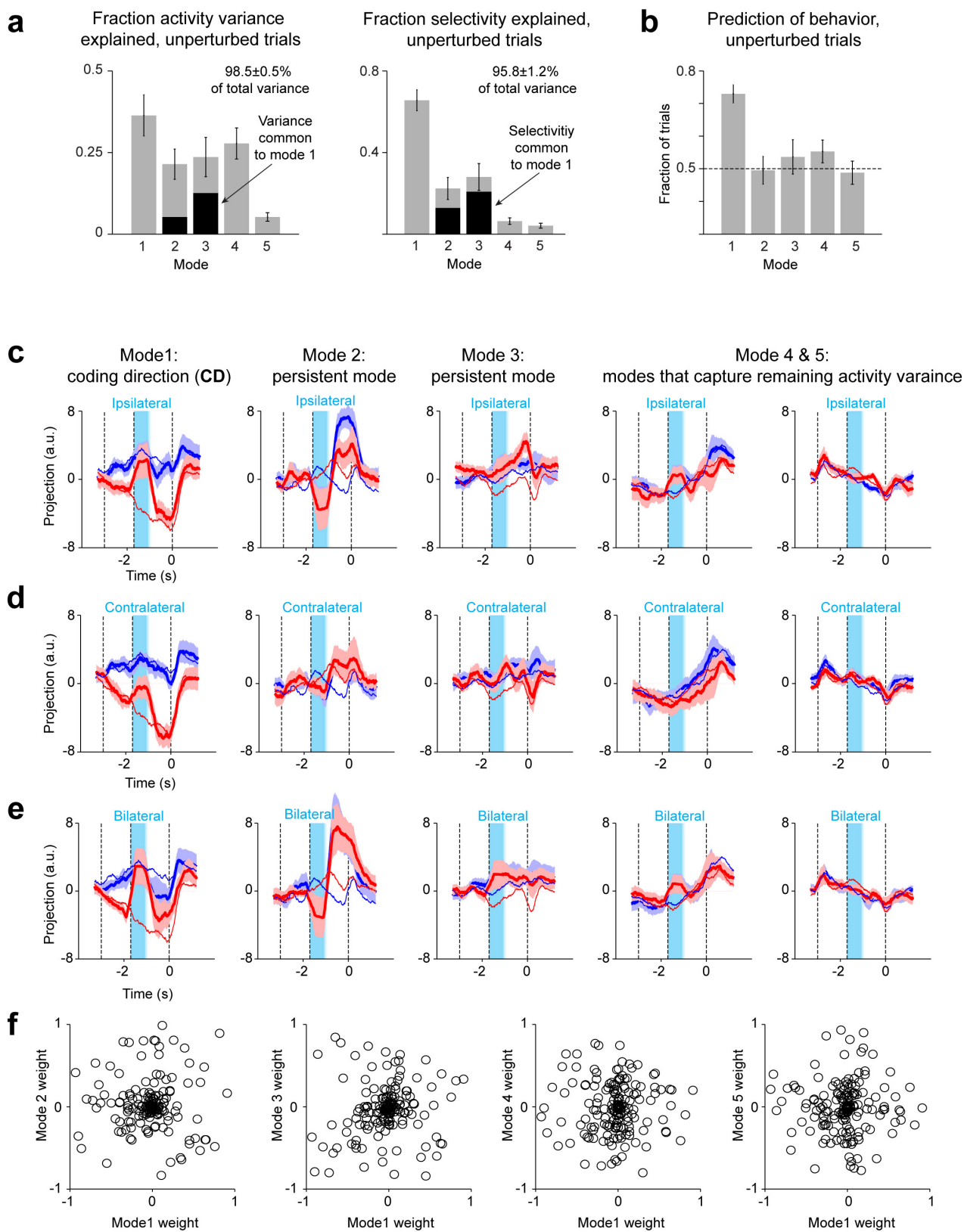
Extended Data Figure 7 | Bilateral photoinhibition disrupts ALM

dynamics and behaviour. **a**, Silicon probe recording during unilateral (4 laser spots) and bilateral (1 laser spot; red box) photoinhibition.

b, Behavioural performance. Bar, mean across all mice ($n = 3$). Symbols, individual mice (mean \pm s.e.m., bootstrap). **c**, Top, significant spike rate changes for individual neurons (black). Neurons (rows) are sorted based on their mean spike rate across the trial epochs. Neurons with mean spike rate below 1 spike s^{-1} or tested for less than 3 trials are excluded ($n = 60$, 59 and 60). Photoinhibition is indicated on the top. Bottom, fraction of neurons with significant spike rate change. **d**, Average population selectivity change from control (Δ selectivity \pm s.e.m. across neurons, bootstrap). Only selective neurons tested for >3 trials in all conditions are shown ($n = 40$). Green lines, time points when the selectivity recovered to 80% of control selectivity (mean \pm s.e.m. across neurons, bootstrap).

Ipsilateral: 490 ± 280 ms to recover to 80% of control selectivity; contralateral: 235 ± 156 ms; bilateral: no recovery at end of delay period.

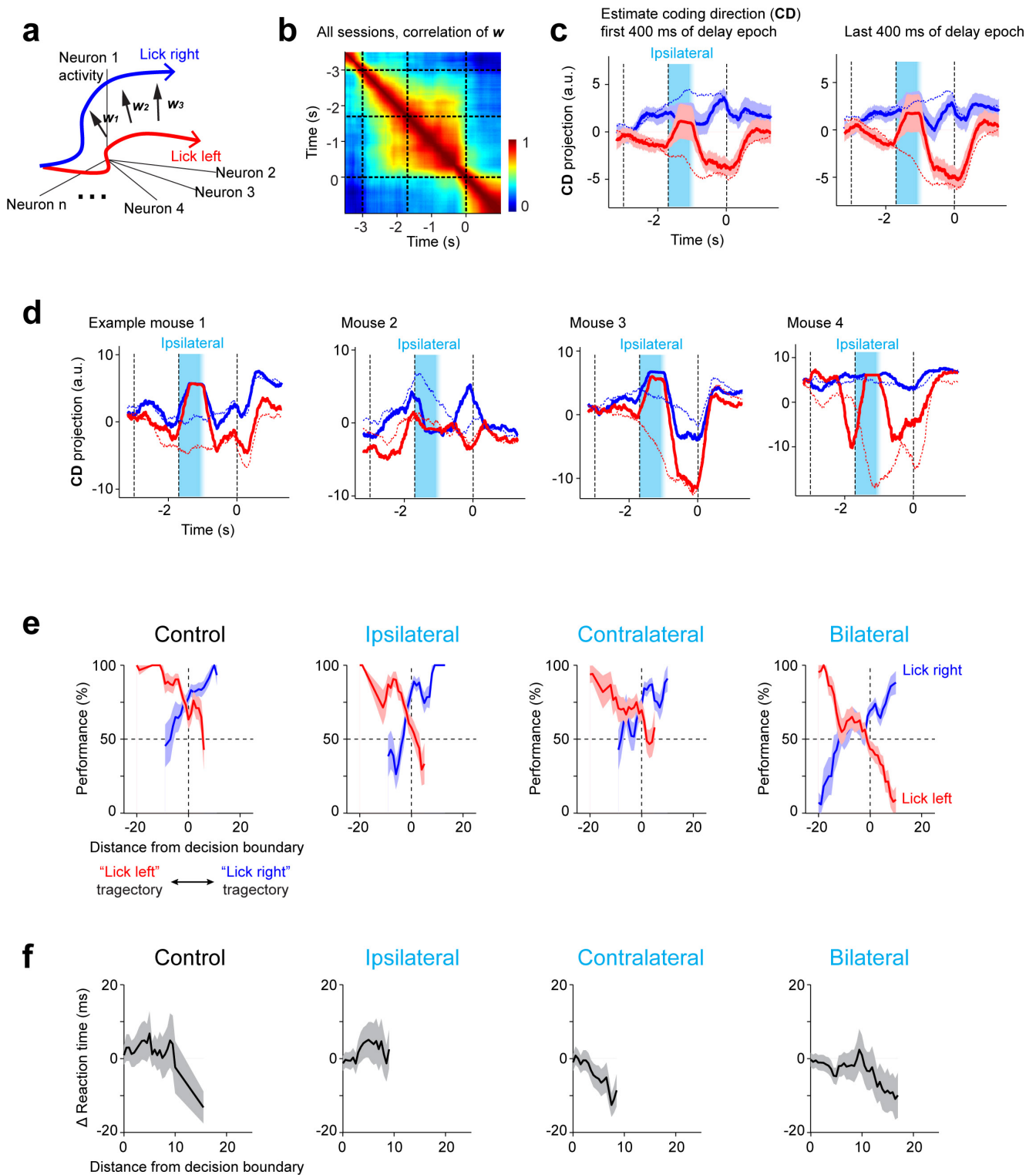
e, Time course of activity trajectories on lick-right (blue) and lick-left (red) trials projected onto the coding direction (CD). Average trajectories from all sessions (\pm s.e.m. across sessions, bootstrap, Methods). From left to right panels: control trials, ipsilateral photoinhibition (4 laser spots), contralateral photoinhibition (4 laser spots), and bilateral photoinhibition (1 laser spot). Dotted line, trajectories in control trials. Only sessions with >5 simultaneously recorded neurons tested for >3 trials in each condition. We quantified the separation between trajectories at the end of delay epoch by computing ROC values for each session: control, 0.80 ± 0.08 ; ipsilateral, 0.64 ± 0.10 ; contralateral, 0.68 ± 0.15 ; bilateral, 0.54 ± 0.8 . Mean \pm s.e.m. across sessions, Methods.



Extended Data Figure 8 | See next page for caption.

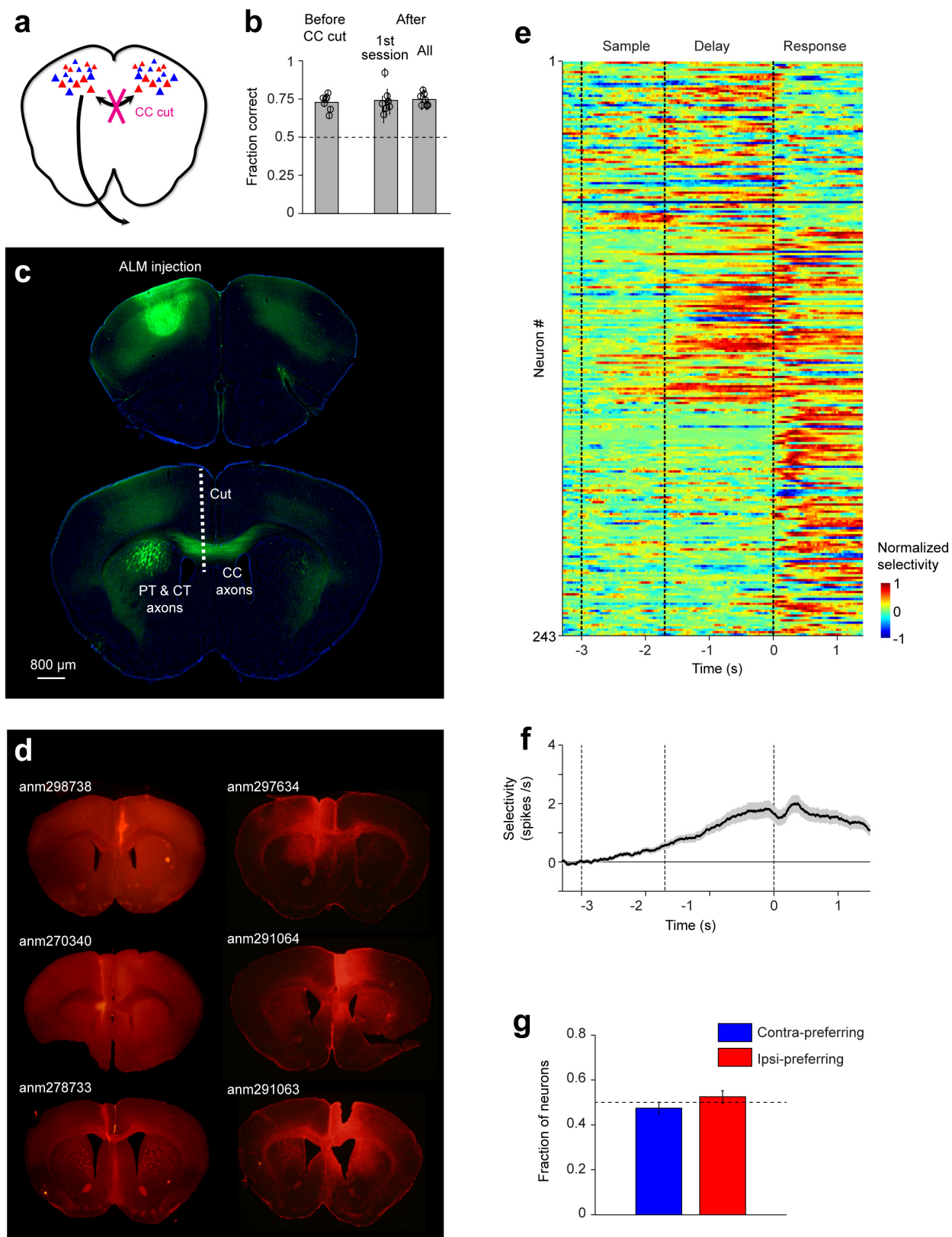
Extended Data Figure 8 | Decomposition of ALM dynamics after perturbation. **a**, Decomposition of activity into five modes based on control trials and ipsilateral perturbations (Methods). Fraction of activity variance (left) and selectivity (right) explained by modes 1–5. The overlap in variance and selectivity between mode 1 and modes 2–3 are highlighted in black. Error bars, s.e.m. across sessions. Data from 16 sessions, 7 mice. Activity variance here is computed using trial-averaged activity (Methods), thus they reflect variance across time and neurons. Activity variance across trials is not reflected. The fraction of variance explained for the single-trial activity would be much lower. **b**, Fraction of upcoming movements predicted based on modes 1–5. Trajectory distance from the decision boundary at the time of the go cue is used to predict behaviour. Lick-right and lick-left trials are pooled. Error bars, s.e.m. across sessions. **c**, Projections of activity along modes 1–5 for ipsilateral perturbation trials (solid). Dashed blue and red lines correspond to the means for control

trials. Error bars, s.e.m. across sessions. For the **CD** mode, a different set of trials was used here to compute **CD** compared to Fig. 3c (Methods). This resulted in small differences in the projected trajectories. **d**, Projections of activity in the same dimensions as in **c** for contralateral perturbation trials. **e**, Projections of activity in the same dimensions as in **c** for bilateral perturbation trials. **f**, Weights of each neuron for mode 1 versus modes 2–5. Mode 1 and modes 2–5 involve overlapping populations of neurons. Data from all sessions were pooled. Note that the ramping modes (4 and 5) are resistant to all perturbations, including bilateral perturbations, suggesting that overall ramping may be driven by a source external to ALM. ROC values between trajectories along the **CD** mode at the end of delay epoch: control, 0.76 ± 0.03 ; ipsilateral, 0.73 ± 0.02 ; contralateral, 0.74 ± 0.03 ; bilateral 0.58 ± 0.03 . ROC values during the time period of photoinhibition: control, 0.72 ± 0.02 ; ipsilateral, 0.54 ± 0.03 ; contralateral, 0.64 ± 0.03 ; bilateral 0.54 ± 0.01 .



Extended Data Figure 9 | ALM dynamics along the coding direction predicts upcoming movements. **a**, Schematic of trajectory analysis in activity space. The difference in the mean response vectors between lick-right and lick-left trials, w , was estimated across different time windows (400 ms) during sample and delay epochs. **b**, w values are similar during sample and delay epoch. Correlation of w values across time. Data from 16 sessions, 7 mice. The coding direction (CD) was taken as the average w value over time. **c**, The recovery of ALM dynamics along the coding direction (CD) is robust to the choice of time window for the calculation of CD. Left, CD was the average w value from the first 400 ms of the delay epoch. Right, CD was the average w value from the last 400 ms of the delay epoch. **d**, The recovery of ALM dynamics along CD is robust across mice.

e, Behavioural performance in lick-right and lick-left trials as a function of trajectory distance from the decision boundary at the time of the go cue. Positive values on the x axis indicate closer distance to the control lick-right trajectory. From left to right panels: control trials, ipsilateral photoinhibition trials, contralateral photoinhibition trials, and bilateral photoinhibition trials. Performance was computed by binning along the CD distance (bin size, 4 on the CD distance scale). s.e.m. was obtained by bootstrapping the trials in each bin. **f**, Reaction times are faster on trials in which the trajectory is far from the decision boundary at the time of the go cue. Δ Reaction time is relative to the mean reaction time from each session. Data from 16 sessions, 7 mice. Data from lick-right and lick-left trials were pooled.



Extended Data Figure 10 | See next page for caption.

Extended Data Figure 10 | Behavioural and ALM dynamics after corpus callosum hemisection. **a**, Schematic. Corpus callosum (CC) was bisected while sparing the pyramidal tract (PT) and corticothalamic (CT) projections. **b**, Behavioural performance. Bar, mean across all mice ($n = 7$). Symbols, individual mice (mean \pm s.e.m., bootstrap). Performance was not affected by the corpus callosum bisection. First session was ~ 17 h after the corpus callosum bisection. **c**, Location of the corpus callosum cut superimposed on axonal projections from ALM. AAV2/1-CAG-EGFP was injected into ALM. A vertical cut ~ 3.5 mm deep was made approximately 0.5 mm from the mid-line. The cut extended from bregma anterior 1.5 mm to posterior 1 mm. The cut was either made in the left hemisphere (3 mice) or the right hemisphere (4 mice). The cut spared the

pyramidal tract and corticothalamic axons. **d**, Coronal section showing the corpus callosum bisection in 6 mice. Left, autofluorescence; right, GFAP immunofluorescence (Methods). **e**, ALM shows normal preparatory activity after the corpus callosum bisection. ALM population selectivity. Selectivity is the difference in spike rate between the preferred and non-preferred trial type, normalized to the peak selectivity (Methods). Only putative pyramidal neurons with significant trial selectivity are shown ($n = 254$ out of 496). In addition, 11 out of 254 neurons tested for < 15 trials for each trial type were excluded. **f**, Average population selectivity in spike rate (black line, \pm s.e.m. across neurons, bootstrap). **g**, Proportion of contra-preferring vs. ipsi-preferring neurons. Error bars, s.e.m. across mice, bootstrap.

Plankton networks driving carbon export in the oligotrophic ocean

Lionel Guidi^{1,2*}, Samuel Chaffron^{3,4,5*}, Lucie Bittner^{6,7,8*}, Damien Eveillard^{9*}, Abdelhalim Larhlimi⁹, Simon Roux^{10†}, Youssef Darzi^{3,4}, Stephane Audic⁸, Léo Berline^{1†}, Jennifer R. Brum^{10†}, Luis Pedro Coelho¹¹, Julio Cesar Ignacio Espinoza¹⁰, Shruti Malviya^{7†}, Shinichi Sunagawa¹¹, Céline Dimier⁸, Stefanie Kandels-Lewis^{11,12}, Marc Picheral¹, Julie Poulain¹³, Sarah Searson^{1,2}, Tara Oceans Consortium Coordinators[‡], Lars Stemmann¹, Fabrice Not⁸, Pascal Hingamp¹⁴, Sabrina Speich¹⁵, Mick Follows¹⁶, Lee Karp-Boss¹⁷, Emmanuel Boss¹⁷, Hiroyuki Ogata¹⁸, Stephane Pesant^{19,20}, Jean Weissenbach^{13,21,22}, Patrick Wincker^{13,21,22}, Silvia G. Acinas²³, Peer Bork^{11,24}, Colomán de Vargas⁸, Daniele Iudicone²⁵, Matthew B. Sullivan^{10†}, Jeroen Raes^{3,4,5}, Eric Karsenti^{7,12}, Chris Bowler⁷ & Gabriel Gorsky¹

The biological carbon pump is the process by which CO₂ is transformed to organic carbon via photosynthesis, exported through sinking particles, and finally sequestered in the deep ocean. While the intensity of the pump correlates with plankton community composition, the underlying ecosystem structure driving the process remains largely uncharacterized. Here we use environmental and metagenomic data gathered during the *Tara Oceans* expedition to improve our understanding of carbon export in the oligotrophic ocean. We show that specific plankton communities, from the surface and deep chlorophyll maximum, correlate with carbon export at 150 m and highlight unexpected taxa such as Radiolaria and alveolate parasites, as well as *Synechococcus* and their phages, as lineages most strongly associated with carbon export in the subtropical, nutrient-depleted, oligotrophic ocean. Additionally, we show that the relative abundance of a few bacterial and viral genes can predict a significant fraction of the variability in carbon export in these regions.

Marine planktonic photosynthetic organisms are responsible for approximately 50% of Earth's primary production and fuel the global ocean biological carbon pump¹. The intensity of the pump is correlated with plankton community composition^{2,3}, and controlled by the relative rates of primary production and carbon remineralization⁴. About 10% of this newly produced organic carbon in the surface ocean is exported through gravitational sinking of particles. Finally, after multiple transformations, a fraction of the exported material reaches the deep ocean where it is sequestered over thousand-year timescales⁵.

Like most biological systems, marine ecosystems in the sunlit upper layer of the ocean (denoted as the euphotic zone) are complex^{6,7}, characterized by a wide range of biotic and abiotic interactions^{8–10} and in constant balance between carbon production, transfer to higher trophic levels, remineralization, and export to the deep layers¹¹. The marine ecosystem structure and its taxonomic and functional composition probably evolved to comply with this loss of energy by modifying organism turnover times and by the establishment of complex

feedbacks between them⁶ and the substrates they can exploit for metabolism¹². Decades of ground-breaking research have focused on identifying independently the key players involved in the biological carbon pump. Among autotrophs, diatoms are commonly attributed to being important in carbon flux because of their large size and fast sinking rates^{13–15}, while small autotrophic picoplankton may contribute directly through subduction of surface water¹⁶ or indirectly by aggregating with larger settling particles or consumption by organisms at higher trophic levels¹⁷. Among heterotrophs, zooplankton such as crustaceans impact carbon flux via production of fast-sinking fecal pellets while migrating hundreds of meters in the water column^{18,19}. These observations, focusing on just a few components of the marine ecosystem, highlight that carbon export results from multiple biotic interactions and that a better understanding of the mechanisms involved in its regulation requires an analysis of the entire planktonic ecosystem.

Advanced sequencing technologies offer the opportunity to simultaneously survey whole planktonic communities and associated

¹Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire d'océanographie de Villefranche (LOV), Observatoire Océanologique, 06230 Villefranche-sur-Mer, France. ²Department of Oceanography, University of Hawaii, Honolulu, Hawaii 96822, USA. ³Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. ⁴Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium. ⁵Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. ⁶Sorbonne Universités, UPMC Univ Paris 06, CNRS, Institut de Biologie Paris-Seine (IBPS), Evolution Paris Seine, F-75005, Paris, France. ⁷Ecole Normale Supérieure, PSL Research University, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS UMR 8197, INSERM U1024, 46 rue d'Ulm, F-75005 Paris, France. ⁸Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire Adaptation et Diversité en Milieu Marin, Station Biologique de Roscoff, 29680 Roscoff, France. ⁹LINA UMR 6241, Université de Nantes, EMN, CNRS, 44322 Nantes, France. ¹⁰Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA. ¹¹Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany. ¹²Directors' Research European Molecular Biology Laboratory Meyerhofstr. 1, 69117 Heidelberg, Germany. ¹³CEA - Institut de Génétique, GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry, France. ¹⁴Aix Marseille Université, CNRS, IGS, UMR 7256, 13288 Marseille, France. ¹⁵Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris CEDEX 05, France. ¹⁶Dept of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. ¹⁷School of Marine Sciences, University of Maine, Orono, Maine 04469, USA. ¹⁸Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan. ¹⁹PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, 28359 Bremen, Germany. ²⁰MARUM, Center for Marine Environmental Sciences, University of Bremen, 28359 Bremen, Germany. ²¹CNRS, UMR 8030, CP 5706 Evry, France. ²²Université d'Evry, UMR 8030, CP 5706 Evry, France. ²³Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM)-CSIC, Pg. Marítim de la Barceloneta 37-49, Barcelona E0800, Spain. ²⁴Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany. ²⁵Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. [†]Present addresses: Department of Microbiology, The Ohio State University, Columbus, Ohio 43210, USA (S.R., J.R.B.); Department of Microbiology, and Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, Ohio 43210, USA (M.B.S.); Aix Marseille Université, CNRS/INSU, Université de Toulon, IRD, Mediterranean Institute of Oceanography (MIO) UM 110, 13288, Marseille, France (L.B.); Biological Oceanography Division, CSIR-National Institute of Oceanography, Dona Paula, Goa 403 004, India (S.M.).

*These authors contributed equally to this work.

‡A list of authors and affiliations appears at the end of the paper.

molecular functions in unprecedented detail. Such a holistic approach may allow the identification of community- or gene-based biomarkers that could be used to monitor and predict ecosystem functions, for example, related to the biogeochemistry of the ocean^{20–22}. Here, we leverage global-scale ocean genomics data sets from the euphotic zone^{10,23–25} and associated environmental data to assess the coupling between ecosystem structure, functional repertoire, and carbon export at 150 m.

Carbon export and plankton community composition

The *Tara* Oceans global circumnavigation crossed diverse ocean ecosystems and sampled plankton at an unprecedented scale^{20,26} (see Methods). Hydrographic data were measured *in situ* or in seawater samples at all stations, as well as nutrients, oxygen and photosynthetic pigments (see Methods). Net primary production (NPP) was derived from satellite measurements (see Methods). In addition, particle size distributions (100 μm to a few millimetres) and concentrations were measured using an underwater vision profiler (UVP) from which carbon export, corresponding to the carbon flux (Fig. 1a) at 150 m, was calculated to range from 0.014 to 18.3 $\text{mg m}^{-2} \text{d}^{-1}$ using methods previously described (see Methods). One should keep in mind that fluxes are calculated from images of particles. These estimates are derived from an approximation of Stokes' law relating the equivalent spherical diameter of particles to carbon flux (see Methods). This exponential approximation is reasonable assuming similar particle composition across all sizes, as highlighted by the standard deviations of parameters in equation (5) (see Methods). Furthermore, because of instrument and method limitations, particles $<250 \mu\text{m}$ were not used, which may underestimate total carbon fluxes. Finally, these fluxes are instantaneous because they do not integrate space and time as sediment traps would. However, the approach allowed us to assemble the largest homogeneous carbon export data set during a single expedition, corresponding to more than 600 profiles over 150 stations. This data set is of similar magnitude to the body of historical data available in the literature that includes the 134 deep sediment trap-based carbon flux time series²⁷ from the JGOFS program and the 419 thorium-derived particulate organic carbon (POC) export measurements²⁸.

From 68 globally distributed sites, a total of 7.2 terabases (Tb) of metagenomics data, representing ~ 40 million non-redundant genes, around 35,000 operational taxonomic units (OTUs) of prokaryotes (Bacteria and Archaea) and numerous mainly uncharacterized viruses and picoeukaryotes, have been described recently^{23,25}. In addition, a set of 2.3 million eukaryotic 18S rDNA ribotypes was generated from a subset of 47 sampling sites corresponding to approximately 130,000 OTUs²⁴. Finally, 5,476 viral 'populations' were identified at 43 sites from viral metagenomic contigs, only 39 ($<0.1\%$) of which had been previously observed²⁵ (see Methods). These genomics data combined across all domains of life and viruses together with carbon export estimates (Fig. 1a) and other environmental parameters were used to explore the relationships between marine biogeochemistry and euphotic plankton communities (see Methods) in the top 150 m of the oligotrophic open ocean. Our study did not include high-latitude areas owing to the current lack of available molecular data and results should not be extrapolated to deeper depths.

Using a method for regression-based modelling of highly multi-dimensional data in biology (specifically a sparse partial least square analysis (sPLS)²⁹, Extended Data Fig. 1), we detected several plankton lineages for which relative sequence abundance correlated with carbon export and other environmental parameters, most notably with NPP, as expected (Fig. 1b and see Supplementary Table 1). These included diatoms, dinoflagellates and Metazoa (zooplankton), lineages classically identified as key contributors to carbon export.

Plankton networks associated with carbon export

While the analysis presented in Fig. 1b supports previous findings about key organisms involved in carbon export from the euphotic

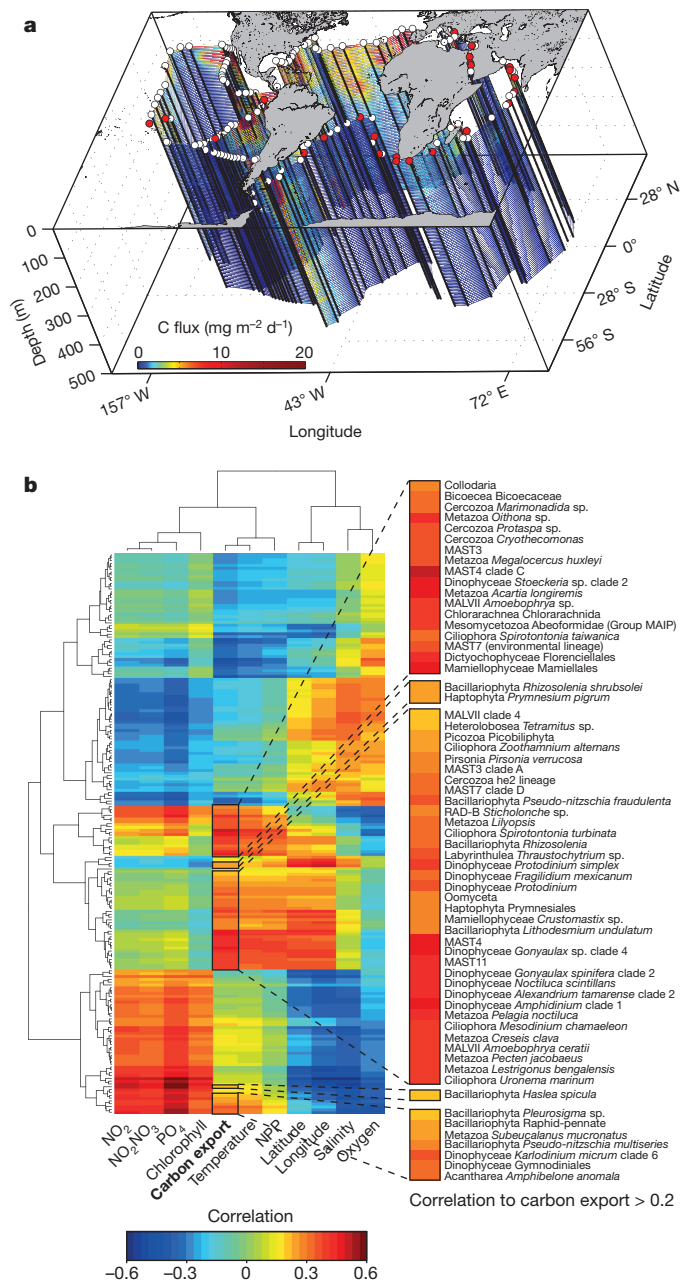


Figure 1 | Global view of carbon fluxes along the *Tara* Oceans circumnavigation route and associated eukaryotic lineages. a, Carbon flux in $\text{mg m}^{-2} \text{d}^{-1}$ and carbon export at 150 m estimated from particle size distribution and abundance measured with the underwater vision profiler (UVP). Stations at which environmental data are available (Supplementary Table 9) are depicted by white dots. Stations at which eukaryotic samples are available are coloured in red (Supplementary Tables 10 and 12). **b**, Eukaryotic lineages associated to carbon export as revealed by standard methods for regression-based modelling (sPLS analysis). Correlations between lineages and environmental parameters are depicted as a clustered heat map and lineages with a correlation to carbon export higher than 0.2 are highlighted (detailed results in Supplementary Table 1).

zone^{14,15,17–19}, it is not able to capture how the intrinsic structure of the planktonic community relates to this biogeochemical process. Conversely, although other recent holistic approaches^{10,30,31} used species co-occurrence networks to reveal potential biotic interactions, they do not provide a robust description of sub-communities driven by abiotic interactions. To overcome these issues, we applied a systems biology approach known as weighted gene correlation network analysis (WGCNA)^{32,33} to detect significant associations between the

Tara Oceans genomics data and carbon export. This method delineates communities in the euphotic zone that are the most associated with carbon export rather than predicting organisms associated with sinking particles.

In brief, the WGCNA approach builds a network in which nodes are features (in this case plankton lineages or gene functions) and links are evaluated by the robustness of co-occurrence scores. WGCNA then clusters the network into modules (hereafter denoted subnetworks) that can be examined to find significant subnetwork–trait relationships. We then filtered each subnetwork using a partial least square (PLS) analysis that emphasizes key nodes (based on the variable importance in projection (VIP) scores; see Methods and Extended Data Fig. 1). These particular nodes are mandatory to summarize a subnetwork (or community) related to carbon export. In particular, they are of interest for evaluating: (i) subnetwork robustness; and (ii) predictive power for a given trait (see Methods and Extended Data Fig. 1).

We applied WGCNA to the relative abundance tables of eukaryotic, prokaryotic and viral lineages^{23–25} and identified unique subnetworks significantly associated with carbon export within each data set (see Methods and Supplementary Tables 2–4). The eukaryotic subnetwork (subnetwork–trait relationship to carbon export, Pearson correlation $r = 0.81$, $P = 5 \times 10^{-15}$) contained 49 lineages (Extended Data Fig. 2a and Supplementary Table 2) among which 20% represented photosynthetic organisms (Fig. 2a and Supplementary Table 2). Surprisingly, this small subnetwork's structure correlates very strongly to carbon export ($r = 0.87$, $P = 5 \times 10^{-16}$, Extended Data Fig. 2d) and it predicts as much as 69% (leave-one-out cross-validated (LOOCV), $R^2 = 0.69$) of the variability in carbon export (Extended Data Fig. 2g). Only ~6% of the subnetwork nodes correspond to diatoms and they show lower VIP scores than dinoflagellates (Supplementary Table 2). This is probably because our samples are not from silicate-replete conditions where diatoms

were blooming. Furthermore, our analysis did not incorporate data from high latitudes, where diatoms are known to be particularly important for carbon export, so this result suggests that dinoflagellates have a heretofore unrecognized role in carbon export processes in subtropical oligotrophic 'type' ecosystems. More precisely, four of the five highest VIP scoring eukaryotic lineages that correlated with carbon export at 150 m were heterotrophs such as Metazoa (copepods), non-photosynthetic Dinophyceae, and Rhizaria (Fig. 2a and Supplementary Table 2). These results corroborate recent metagenomics analysis of microbial communities from sediment traps in the oligotrophic North Pacific subtropical gyre³⁴. Consistently, *in situ* imaging surveys have revealed Rhizarian lineages, made up of large fragile organisms such as the Collodaria, to represent an until now under-appreciated component of global plankton biomass (T. Biard *et al.*, submitted), which here also appear to be of relevance for carbon export. Another 14% of lineages from the subnetwork correspond to parasitic organisms, a largely unexplored component of planktonic ecosystems when studying carbon export.

The prokaryotic subnetwork that associated most significantly with carbon export at 150 m (subnetwork–trait relationship to carbon export, $r = 0.32$, $P = 9 \times 10^{-3}$) contained 109 OTUs (Extended Data Fig. 2b and Supplementary Table 3), its structure correlated well to carbon export ($r = 0.47$, $P = 5 \times 10^{-6}$, Extended Data Fig. 2e) and it could predict as much as 60% of the carbon export variability (LOOCV, $R^2 = 0.60$) (Extended Data Fig. 2h). By far the highest VIP score within this community was assigned to *Synechococcus*, followed by *Cobetia*, *Pseudoalteromonas* and *Idiomarina*, as well as *Vibrio* and *Arcobacter* (Fig. 2b and Supplementary Table 3). Noteworthy, the genus *Prochlorococcus* and SAR11 clade fall out of this community, while the significance of *Synechococcus* for carbon export could be validated using absolute cell counts estimated by flow cytometry ($r = 0.64$,

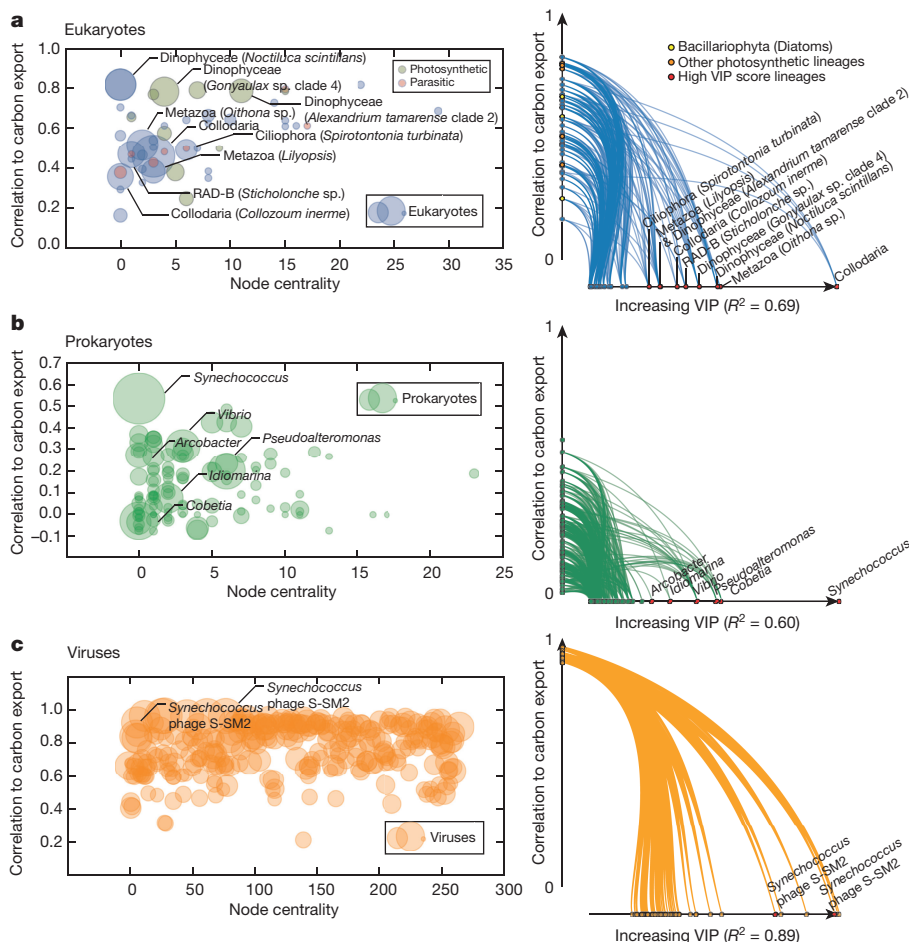


Figure 2 | Ecological networks reveal key lineages associated with carbon export at 150 m at global scale. The relative abundances of taxa in selected subnetworks were used to estimate carbon export and to identify key lineages associated with the process. **a**, The selected eukaryotic subnetwork ($n = 49$, see Supplementary Table 2) can predict carbon export with high accuracy (PLS regression, LOOCV, $R^2 = 0.69$, see Extended Data Fig. 2g). Lineages with the highest VIP score (dot size is proportional to the VIP score in the scatter plot) in the PLS are depicted as red dots corresponding to three Rhizaria (Collodaria, *Collozoum inerme* and *Sticholonche* sp.), one copepod (*Oithona* sp.), one siphonophore (*Lilyopsis*), three Dinophyceae and one ciliate (*Spirotonion turbinata*). **b**, The selected prokaryotic subnetwork ($n = 109$, see Supplementary Table 3) can predict carbon export with good accuracy (PLS regression, LOOCV, $R^2 = 0.60$, see Extended Data Fig. 2h). The selected viral population subnetwork ($n = 277$, see Supplementary Table 4) can predict carbon export with high accuracy (PLS regression, LOOCV, $R^2 = 0.89$, see Extended Data Fig. 2i). Two viral populations with a high VIP score (red dots) are predicted as *Synechococcus* phages (see Supplementary Table 4).

$P = 4 \times 10^{-10}$, Extended Data Fig. 2k). Moreover, *Prochlorococcus* cell counts did not correlate with carbon export ($r = -0.13$, $P = 0.27$, Extended Data Fig. 2j) whereas the *Synechococcus* to *Prochlorococcus* cell count ratio correlated positively and significantly ($r = 0.54$, $P = 4 \times 10^{-7}$, Extended Data Fig. 2l), suggesting the relevance of *Synechococcus*, rather than *Prochlorococcus*, to carbon export. Notably, *Pseudoalteromonas*, *Idiomarina*, *Vibrio* and *Arcobacter* (of which several species are known to be associated with eukaryotes³⁵) have also been observed in live and poisoned sediment traps³⁴ and display very high VIP scores in the subnetwork associated with carbon export. Additional genera reported as being enriched in poisoned traps (also known as being associated with eukaryotes) include *Enterovibrio* and *Campylobacter*, and are present as well in the carbon export associated subnetwork.

Interestingly, the viral subnetwork (involving 277 populations) most related to carbon export at 150 m ($r = 0.93$, $P = 2 \times 10^{-15}$, Extended Data Fig. 2c) contained particularly high VIP scores for two *Synechococcus* phages (Fig. 2c and Supplementary Table 4), which represented a 16-fold enrichment (Fisher's exact test $P = 6.4 \times 10^{-9}$). Its structure also correlated with carbon export ($r = 0.88$, $P = 6 \times 10^{-93}$, Extended Data Fig. 2f) and could predict up to 89% of the variability of carbon export (LOOCV, $R^2 = 0.89$) (Extended Data Fig. 2i). The significance of these convergent results is reinforced by the fact that sequences from these data sets are derived from organisms collected on distinct filters with different mesh sizes (see Methods), and further implicates the importance of top-down processes in carbon export.

With the aim of integrating eukaryotic, prokaryotic, and viral communities in the euphotic zone with carbon export at 150 m, we synthesized their respective subnetworks using a single global co-occurrence network established previously¹⁰. The resulting network focused on key lineages and their predicted co-occurrences (Fig. 3). Lineages with high VIP values (such as *Synechococcus*) are revealed as hubs of the co-occurrence network¹⁰, illustrating the potentially strategic key roles within the integrated network of lineages under-appreciated by conventional methods to study carbon export. Associations between the hub lineages are mostly mutually exclusive, which may explain the relatively

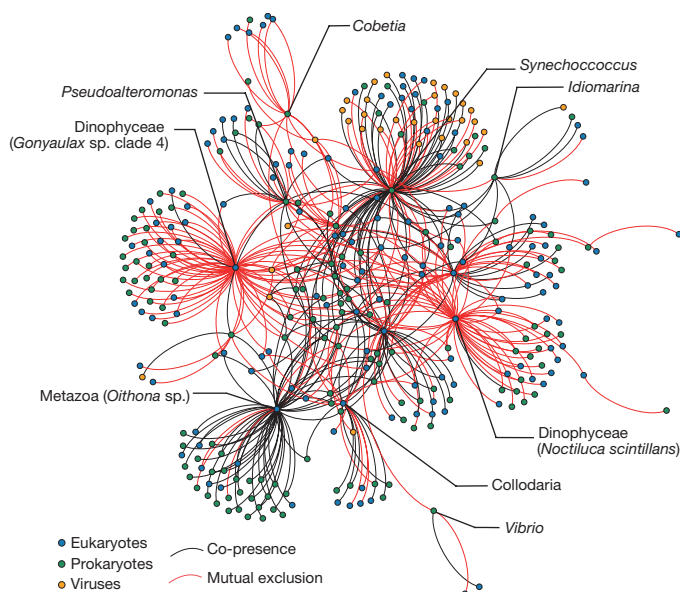


Figure 3 | Integrated plankton community network built from eukaryotic, prokaryotic and viral subnetworks related to carbon export at 150 m. Major lineages were selected within the three subnetworks (VIP > 1) (Supplementary Tables 2, 3 and 4). Co-occurrences between all lineages of interest were extracted, if present, from a previously established global co-occurrence network (see Methods). Only lineages discussed within the study are pinpointed. The resulting graph is composed of 329 nodes, 467 edges, with a diameter of 7, and average weighted degree of 4.6.

weak correlation of some of these lineages with carbon export when using standard correlation analyses, as shown in Fig. 1b.

Gene functions associated with carbon export

Given the potential importance of prokaryotic processes influencing the biological carbon pump²², we used the same analytical approaches to examine the prokaryotic genomic functions associated with carbon export at 150 m in the annotated Ocean Microbial Reference Gene Catalogue from Tara Oceans²³. We built a global co-occurrence network for functions (that is, orthologous groups of genes (OGs)) from the euphotic zone and identified two subnetworks of functions that are significantly associated with carbon export (light and dark green subnetworks; FNET1 and FNET2, respectively, see Extended Data Fig. 3a–c).

The majority of functions in FNET1 and FNET2 correlate well with carbon export (FNET1: mean $r = 0.45$, s.d. = 0.09 and FNET2: mean $r = 0.34$, s.d. = 0.10). Interestingly, FNET2 functions ($n = 220$) encode mostly (83%) core functions (that is, functions observed in all euphotic samples, see Methods) while the majority of FNET1 functions ($n = 441$) are non-core (85%) (see Supplementary Tables 5 and 6), highlighting both essential and adaptive ecological functions associated with carbon export. Top VIP scoring functions in the FNET1 subnetwork are membrane proteins such as ABC-type sugar transporters (Extended Data Fig. 3c). This subnetwork also contains many functions specific to the *Synechococcus* accessory photosynthetic apparatus (for example, relating to phycobilisomes, phycocyanin and phycoerythrin; see Supplementary Table 5), which is consistent with the major role of this genus for carbon export inferred from the prokaryotic subnetwork (Fig. 2b). In addition, functions related to carbohydrates, inorganic ion transport and metabolism, as well as transcription, are also well represented (Fig. 4), suggesting overall a subnetwork of functions dedicated to photosynthesis and growth.

The FNET2 subnetwork contains several functions encoded by genes taxonomically assigned to *Candidatus pelagibacter* and *Prochlorococcus*, known as occupying similar oceanic regions as *Synechococcus*, but overall most of its relative abundance (74%) is taxonomically unclassified (Extended Data Fig. 3e). Top VIP scoring functions in FNET2 are also membrane proteins and ABC-type sugar transporters, as well as functions involved in carbohydrate breakdown such as a chitinase (Extended Data Fig. 3c). These features highlight the potential roles of bacteria in the formation and degradation of marine aggregates³⁶. Notably, 77% and 58%, of OGs with a VIP score > 1 in FNET1 and FNET2, respectively, are functionally uncharacterized^{37,38} (Fig. 4), pointing to the strong need for future molecular work to explore these functions (see Supplementary Tables 5 and 6).

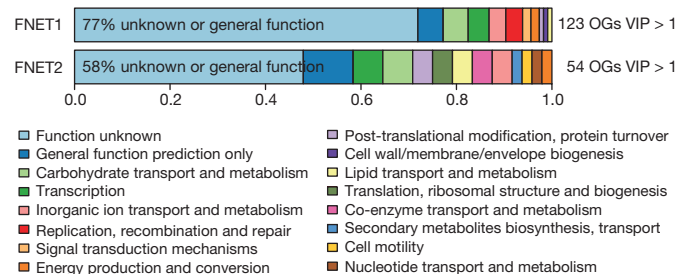


Figure 4 | Key bacterial functional categories associated with carbon export at 150 m at global scale. A bacterial functional network was built based on orthologous group/gene (OG) relative abundances using the WGCNA methodology (see Methods) and correlated to classical oceanographic parameters. Two functional subnetworks (FNET1 ($n = 220$) and FNET2 ($n = 441$), respectively, Extended Data Fig. 3a) are significantly associated with carbon export (FNET1: $r = 0.42$, $P = 4 \times 10^{-9}$ and FNET2: $r = 0.54$, $P = 7 \times 10^{-6}$, see Extended Data Fig. 3b). Higher functional categories are depicted for functions with a VIP score > 1 (PLS regression, LOOCV, FNET1 $R^2 = 0.41$ and FNET2 $R^2 = 0.48$, see Extended Data Fig. 3d) in both subnetworks.

As for plankton communities, the relevance of the identified bacterial functions to predict carbon export was also confirmed by PLS regression (Extended Data Fig. 3d). The functional subnetworks predict 41% and 48% of carbon export variability (LOOCV, $R^2 = 0.41$ and 0.48 for FNET1 and FNET2, respectively) with a minimal number of functions (Fig. 4, 123 and 54 functions with a VIP score >1 for FNET1 and FNET2, respectively). Finally, higher predictive power was obtained using subnetworks of viral protein clusters (Extended Data Fig. 4a–c), predicting 55% and 89% of carbon export variability (LOOCV $R^2 = 0.55$ and 0.89 for VNET1 and VNET2, respectively; Extended Data Fig. 4d, Supplementary Tables 7 and 8), suggesting a key role of not only bacteria, but also their phages in processes sustaining carbon export at a global level.

Discussion

In this work we reveal the potential contribution of unexpected components of plankton communities, and confirm the importance of prokaryotes and viruses for carbon export in the nutrient-depleted oligotrophic ocean. Carbon export at 150 m has been estimated from particle size distribution in a global data set, but should be taken with caution, as the estimates do not account for particle composition. In addition, these export estimates evaluate how much carbon leaves the euphotic zone, but they are not related and should not be extrapolated to sequestration, which occurs after remineralization, deeper in the water column, and over longer timescales. Nonetheless, the use of the UVP was the only realistic method to evaluate carbon flux over the 3-year expedition because deployment of sediment traps at all stations would have been impossible. While our findings are consistent with the numerous previous studies that have highlighted the central role of copepods and diatoms in carbon export^{14,15,17–19}, they place them in an ecosystem context and reveal hypothetical processes correlating with the intensity of export, such as parasitism, infection and predation. For example, while viruses are commonly assumed to lyse cells and maintain fixed organic carbon in surface waters, thereby reducing the intensity of the biological carbon pump³⁹, there are hints that viral lysis may increase carbon export through the production of colloidal particles and aggregate formation⁴⁰. Our current study suggests that these latter roles may be more ubiquitous than currently appreciated. The importance of aggregation and cell stickiness as inferred from gene network analysis should be further explored mechanistically to investigate the biological significance of these findings.

The future evolution of the oceanic carbon sink remains uncertain because of poorly constrained processes, particularly those associated with the biological pump. With current trends in climate change, the size and biodiversity of phytoplankton are predicted to decrease globally^{41,42}. Furthermore, in spite of the potential importance of viruses revealed in this study, they have largely been ignored because of limitations in sampling technologies. Consequently, as oligotrophic gyres expand and global mean NPP decreases⁴³, the field is currently unable to predict the consequences for carbon export from the ocean's euphotic zone. By pinpointing key lineages and key microbial functions that correlate with carbon export at 150 m in these areas, this study provides a framework to address this critical bottleneck. However, the associations presented do not necessarily suggest a causal effect on carbon export, which will require further investigation.

One of the grand challenges in the life sciences is to link genes to ecosystems⁴⁴, based on the posit that genes can have predictable ecological footprints at community and ecosystem levels^{45–47}. The Tara Oceans data sets have allowed us to predict as much as 89% of the variability in carbon export from the oligotrophic surface ocean with just a small number of genes, largely with unknown functions, encoded by prokaryotes and viruses. These findings can be used as a basis to include biological complexity and guide experimental work designed to inform climate modelling of the global carbon cycle. Such statistical analyses, scaling from genes to ecosystems, may open the way to

the development of a new conceptual and methodological framework to better understand the mechanisms underpinning key ecological processes.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 May; accepted 18 December 2015.

Published online 10 February 2016.

- Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281**, 237–240 (1998).
- Boyd, P. W. & Newton, P. Evidence of the potential influence of planktonic community structure on the interannual variability of particulate organic-carbon flux. *Deep Sea Res. Part I Oceanogr. Res. Pap.* **42**, 619–639 (1995).
- Guidi, L. *et al.* Effects of phytoplankton community on production, size, and export of large aggregates: a world-ocean analysis. *Limnol. Oceanogr.* **54**, 1951–1963 (2009).
- Kwon, E. Y., Primeau, F. & Sarmiento, J. L. The impact of remineralization depth on the air-sea carbon balance. *Nature Geosci.* **2**, 630–635 (2009).
- IPCC. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* (Cambridge University Press, 2013).
- Kitano, H. Biological robustness. *Nature Rev. Genet.* **5**, 826–837 (2004).
- Suweis, S., Simini, F., Banavar, J. R. & Maritan, A. Emergence of structural and dynamical properties of ecological mutualistic networks. *Nature* **500**, 449–452 (2013).
- Chow, C. E. T., Kim, D. Y., Sachdeva, R., Caron, D. A. & Fuhrman, J. A. Top-down controls on bacterial community structure: microbial network analysis of bacteria, T4-like viruses and protists. *ISME J.* **8**, 816–829 (2014).
- Fuhrman, J. A. Microbial community structure and its functional implications. *Nature* **459**, 193–199 (2009).
- Lima-Mendez, G. *et al.* Determinants of community structure in the global plankton interactome. *Science* **348**, (2015).
- Giering, S. L. C. *et al.* Reconciliation of the carbon budget in the ocean's twilight zone. *Nature* **507**, 480–483 (2014).
- Azam, F. Microbial control of oceanic carbon flux: the plot thickens. *Science* **280**, 694–696 (1998).
- Agusti, S. *et al.* Ubiquitous healthy diatoms in the deep sea confirm deep carbon injection by the biological pump. *Nature Commun.* **6**, 7608 (2015).
- Sancetta, C., Villareal, T. & Falkowski, P. Massive fluxes of rhizosolenid diatoms – a common occurrence. *Limnol. Oceanogr.* **36**, 1452–1457 (1991).
- Scharek, R., Tupas, L. M. & Karl, D. M. Diatom fluxes to the deep sea in the oligotrophic north Pacific gyre at station ALOHA. *Mar. Ecol. Prog. Ser.* **182**, 55–67 (1999).
- Omand, M. M. *et al.* Eddy-driven subduction exports particulate organic carbon from the spring bloom. *Science* **348**, 222–225 (2015).
- Richardson, T. L. & Jackson, G. A. Small phytoplankton and carbon export from the surface ocean. *Science* **315**, 838–840 (2007).
- Steinberg, D. K. *et al.* Bacterial vs. zooplankton control of sinking particle flux in the ocean's twilight zone. *Limnol. Oceanogr.* **53**, 1327–1338 (2008).
- Turner, J. T. Zooplankton fecal pellets, marine snow, phytodetritus and the ocean's biological pump. *Prog. Oceanogr.* **130**, 205–248 (2015).
- Karsenti, E. *et al.* A holistic approach to marine eco-systems biology. *PLoS Biol.* **9**, (2011).
- Strom, S. L. Microbial ecology of ocean biogeochemistry: a community perspective. *Science* **320**, 1043–1045 (2008).
- Worden, A. Z. *et al.* Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. *Science* **347**, 1257594 (2015).
- Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
- de Vargas, C. *et al.* Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605 (2015).
- Brum, J. R. *et al.* Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498 (2015).
- Bork, P. *et al.* Tara Oceans studies plankton at planetary scale. *Science* **348**, 873 (2015).
- Honjo, S., Manganini, S. J., Krishfield, R. A. & Francois, R. Particulate organic carbon fluxes to the ocean interior and factors controlling the biological pump: A synthesis of global sediment trap programs since 1983. *Prog. Oceanogr.* **76**, 217–285 (2008).
- Henson, S. A., Sanders, R. & Madsen, E. Global patterns in efficiency of particulate organic carbon export and transfer to the deep ocean. *Glob. Biogeochem. Cycles* **26**, (2012).
- Lê Cao, K. A., Rossouw, D., Robert-Granié, C. & Besse, P. A sparse PLS for variable selection when integrating omics data. *Stat. Appl. Genet. Mol. Biol.* **7**, 35 (2008).
- Chaffron, S., Rehrauer, H., Pernthaler, J. & von Mering, C. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res.* **20**, 947–959 (2010).

31. Faust, K. & Raes, J. Microbial interactions: from networks to models. *Nature Rev. Microbiol.* **10**, 538–550 (2012).
32. Aylward, F. O. *et al.* Microbial community transcriptional networks are conserved in three domains at ocean basin scales. *Proc. Natl Acad. Sci.* **112**, 5443–5448 (2015).
33. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9** (2008).
34. Fontanez, K. M., Eppley, J. M., Samo, T. J., Karl, D. M. & DeLong, E. F. Microbial community structure and function on sinking particles in the North Pacific Subtropical Gyre. *Front. Microbiol.* **6**, (2015).
35. Thomas, T. *et al.* Analysis of the *Pseudoalteromonas tunicata* genome reveals properties of a surface-associated life style in the marine environment. *PLoS ONE* **3**, (2008).
36. Azam, F. & Malfatti, F. Microbial structuring of marine ecosystems. *Nature Rev. Microbiol.* **5**, 782–791 (2007).
37. Shi, Y., Tyson, G. W. & DeLong, E. F. Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* **459**, 266–269 (2009).
38. Yooshef, S. *et al.* The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol.* **5**, e16 (2007).
39. Suttle, C. A. Marine viruses – major players in the global ecosystem. *Nature Rev. Microbiol.* **5**, 801–812 (2007).
40. Weinbauer, M. G. Ecology of prokaryotic viruses. *FEMS Microbiol. Rev.* **28**, 127–181 (2004).
41. Finkel, Z. V. *et al.* Phytoplankton in a changing world: cell size and elemental stoichiometry. *J. Plankton Res.* **32**, 119–137 (2010).
42. Sommer, U. & Lewandowska, A. Climate change and the phytoplankton spring bloom: warming and overwintering zooplankton have similar effects on phytoplankton. *Glob. Change Biol.* **17**, 154–162 (2011).
43. Behrenfeld, M. J. *et al.* Climate-driven trends in contemporary ocean productivity. *Nature* **444**, 752–755 (2006).
44. DeLong, E. F. *et al.* Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**, 496–503 (2006).
45. Gianoulis, T. A. *et al.* Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc. Natl Acad. Sci. USA* **106**, 1374–1379 (2009).
46. Tilman, D. *et al.* The influence of functional diversity and composition on ecosystem processes. *Science* **277**, 1300–1302 (1997).
47. Wymore, A. S. *et al.* Genes to ecosystems: exploring the frontiers of ecology with one of the smallest biological units. *New Phytol.* **191**, 19–36 (2011).
48. Picheral, M. *et al.* Vertical profiles of environmental parameters measured on discrete water samples collected with Niskin bottles during the Tara Oceans expedition 2009–2013. *PANGAEA* <http://dx.doi.org/10.1594/PANGAEA.836319> (2014).
49. Picheral, M. *et al.* Vertical profiles of environmental parameters measured from physical, optical and imaging sensors during Tara Oceans expedition 2009–2013. *PANGAEA* <http://dx.doi.org/10.1594/PANGAEA.836321> (2014).
50. Chaffron, S. *et al.* Contextual environmental data of selected samples from the Tara Oceans Expedition (2009–2013). *PANGAEA* <http://dx.doi.org/10.1594/PANGAEA.840718> (2014).
51. Pesant, S. *et al.* Open science resources for the discovery and analysis of Tara Oceans data. *Sci. Data* **2**, 150023 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the commitment of the following people and sponsors: CNRS (in particular Groupement de Recherche GDR3280), European Molecular Biology Laboratory (EMBL), Genoscope/CEA, VIB, Stazione Zoologica Anton Dohrn, UNIMIB, Fund for Scientific Research – Flanders, Rega Institute, KU Leuven, The French Ministry of Research, the French Government 'Investissements d'Avenir' programmes OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), MEMO LIFE (ANR-10-LABX-54), PSL* Research University (ANR-11-IDEX-0001-02), ANR (projects POSEIDON/ANR-09-BLAN-0348, PHYTBACK/ANR-2010-1709-01, PROMETHEUS/ANR-09-PCS-GENM-217, TARA-GIRUS/ANR-09-PCS-GENM-218, SAMOSA, ANR-13-ADAP-0010), European Union FP7 (MicroB3/No.287589, ERC Advanced Grant Award to C.B. (Diatomite: 294823), Gordon and Betty Moore Foundation grant (#3790 and #2631) and the UA Technology and Research Initiative Fund and the Water, Environmental, and Energy Solutions Initiative to M.B.S., the Italian Flagship Program RITMARE to D.L., the Spanish Ministry of Science and Innovation grant CGL2011-26848/BOS MicroOcean PANGENOMICS to S.G.A., TANIT (CONES 2010-0036) from the Agència de Gestió d'Ajuts Universitaris i Reserca to S.G.A., JSPS KAKENHI grant number 26430184 to H.O., and FWO, BIO5, Biosphere 2 to M.B.S. We also thank the support and commitment of Agnès B. and Etienne Bourgois, the Veolia Environment Foundation, Région Bretagne, Lorient Agglomération, World Courier, Illumina, the EDF Foundation, FRB, the Prince Albert II de Monaco Foundation, the Tara schooner and its captains and crew. We thank MERCATOR-CORIOLIS and ACRI-ST for providing daily satellite data during the expedition. We are also grateful to the French Ministry of Foreign Affairs for supporting the expedition and to the countries

who graciously granted sampling permissions. Tara Oceans would not exist without continuous support from 23 institutes (<http://oceans.taraexpeditions.org>). The authors further declare that all data reported herein are fully and freely available from the date of publication, with no restrictions, and that all of the samples, analyses, publications, and ownership of data are free from legal entanglement or restriction of any sort by the various nations whose waters the Tara Oceans expedition sampled in. This article is contribution number 34 of Tara Oceans.

Author Contributions L.G., S.C., Lu.B. and D.E. designed the study and wrote the paper. C.D., M.P., J.P. and Sa.S. collected Tara Oceans samples. S.K.-L. managed the logistics of the Tara Oceans project. L.G. and M.P. analysed oceanographic data. S.C. and Lu.B. analysed taxonomic data. S.C., Lu.B., D.E. and S.R. performed the genomic and statistical analyses. A.L., Y.D., L.G., S.C., Lu.B. and D.E. produced and analysed the networks. E.K., C.B. and G.G. supervised the study. M.S., J.R., E.K., C.B. and G.G. provided constructive comments, revised and edited the manuscript. Tara Oceans coordinators provided constructive criticism throughout the study. All authors discussed the results and commented on the manuscript.

Author Information Data described herein is available at European Nucleotide Archive under the project identifiers PRJEB402, PRJEB6610 and PRJEB7988, PANGAEA^{48–50}, and a companion website (<http://www.raeslab.org/companion/ocean-carbon-export.html>). The data release policy regarding future public release of Tara Oceans data is described in ref. 51. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.G. (lguidi@obs-vlfr.fr), S.C. (samuel.chaffron@vib-kuleuven.be), Lu.B. (lucie.bittner@upmc.fr), D.E. (damien.eveillard@univ-nantes.fr), J.R. (Jeroen.Raes@vib-kuleuven.be), E.K. (karsenti@embl.de), C.B. (cbowler@biologie.ens.fr) or G.G. (gorsky@obs-vlfr.fr).

Tara Oceans Consortium Coordinators

Silvia G. Acinas¹, Peer Bork^{2,3}, Emmanuel Boss⁴, Chris Bowler⁵, Colomán de Vargas⁶, Michael Follows⁷, Gabriel Gorsky⁸, Nigel Grimsley⁹, Pascal Hingamp¹⁰, Daniele Iudicone¹¹, Olivier Jaillon^{12,13,14}, Stefanie Kandels-Lewis^{15,16}, Lee Karp-Boss⁴, Eric Karsenti^{5,16}, Fabrice Noté⁶, Hiroyuki Ogata¹⁷, Stéphane Pesant^{18,19}, Jeroen Raes^{20,21,22}, Christian Sardet²³, Mike Sieracki²⁴, Sabrina Speich²⁵, Lars Stemmann⁸, Matthew B. Sullivan^{26†}, Shinichi Sunagawa¹⁵, Patrick Wincker^{12,13,14}

¹Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM)-CSIC, Pg. Marítim de la Barceloneta 37-49, Barcelona E0800, Spain. ²Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany.

³Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany. ⁴School of Marine Sciences, University of Maine, Orono, Maine 04469, USA. ⁵Ecole Normale Supérieure, PSL

Research University, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS UMR

8197, INSERM U1024, 46 rue d'Ulm, F-75005 Paris, France. ⁶Sorbonne Universités, UPMC

Université Paris 06, CNRS, Laboratoire Adaptation et Diversité en Milieu Marin, Station

Biologique de Roscoff, 29680 Roscoff, France. ⁷Department of Earth, Atmospheric and

Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139,

USA. ⁸Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire d'Océanographie

de Villefranche (LOV), Observatoire Océanologique, 06230 Villefranche-sur-Mer, France.

⁹Sorbonne Universités, UPMC Université Paris 06, CNRS, Biologie Intégrative des Organismes

Marins (BIOM), Observatoire Océanologique de Banyuls, 66650 Banyuls-sur-Mer France,

France. ¹⁰Aix Marseille Université, CNRS, IGS, UMR 7256, 13288 Marseille, France. ¹¹Stazione

Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. ¹²CEA - Institut de Génétique,

GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry, France. ¹³CNRS, UMR 8030, CP5706 Evry,

France. ¹⁴Université d'Evry, UMR 8030, CP5706 Evry, France. ¹⁵Structural and Computational

Biology, European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany.

¹⁶Directors' Research European Molecular Biology Laboratory Meyerhofstr. 1, 69117

Heidelberg, Germany. ¹⁷Institute for Chemical Research, Kyoto University, Gokasho, Uji,

Kyoto, 611-0011, Japan. ¹⁸PANGAEA, Data Publisher for Earth and Environmental Science,

University of Bremen, 28359 Bremen, Germany. ¹⁹MARUM, Center for Marine Environmental

Sciences, University of Bremen, 28359 Bremen, Germany. ²⁰Department of Microbiology and

Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. ²¹Center for

the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium. ²²Department of Applied

Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. ²³Sorbonne

Universités, UPMC Université Paris 06, CNRS, Laboratoire de biologie du développement

(LBDV), Observatoire Océanologique, 06230 Villefranche-sur-Mer, France. ²⁴Bigelow Laboratory

for Ocean Science, East Boothbay ME 04544, USA. ²⁵Department of Geosciences, Laboratoire

de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris

CEDEX 05, France. ²⁶Department of Ecology and Evolutionary Biology, University of Arizona,

Tucson, Arizona 85721, USA.

†Present addresses: National Science Foundation, Arlington, 22230 Virginia, USA (M.S.);

Department of Microbiology, and Department of Civil, Environmental and Geodetic Engineering,

The Ohio State University, Columbus, Ohio 43210, USA (M.B.S.).

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Environmental data collection. From 2009–2013, environmental data (Supplementary Table 9) were collected across all major oligotrophic oceanic provinces in the context of the *Tara* Oceans expeditions²⁰. Sampling stations were selected to represent distinct marine ecosystems at a global scale⁵¹. Note that Southern Ocean stations were not examined herein because they were ranked as outliers due to their exceptional environmental characteristics and biota^{23,24}. Environmental data were obtained from vertical profiles of a sampling package^{48,49}. It consisted of conductivity and temperature sensors, chlorophyll and CDOM fluorometers, light transmissometer (Wetlabs C-star 25 cm), a backscatter sensor (WetLabs ECO BB), a nitrate sensor (SATLANTIC ISUS) and an underwater vision profiler (Hydroptics UVP⁵²). Nitrate and fluorescence to chlorophyll concentrations as well as salinity were calibrated with water samples collected with Niskin bottle⁴⁸. Net primary production (NPP) data were extracted from 8-day composites of the vertically generalized production model (VGPM)⁵³ at the week of sampling⁵⁰. Carbon fluxes and carbon export, corresponding to the carbon flux at 150 m, were estimated based on particle concentration and size distributions obtained from the UVP⁴⁹ and details are presented below.

From particle size distribution to carbon export estimation. Previous research has shown that the distribution of particle size follows a power law over the micrometre to the millimetre size range^{3,54,55}. This Junge-type distribution translates into the following mathematical equation, whose parameters can be retrieved from UVP images:

$$n(d) = ad^k \quad (1)$$

where d is the particle diameter, and exponent k is defined as the slope of the number spectrum when equation (1) is log transformed. This slope is commonly used as a descriptor of the shape of the aggregate size distribution.

The carbon-based particle size approach relies on the assumption that the total carbon flux of particles (F) corresponds to the flux spectrum integrated over all particle sizes:

$$F = \int_0^\infty n(d)m(d)w(d)dd \quad (2)$$

where $n(d)$ is the particle size spectrum, that is, equation (1), and $m(d)$ is the mass (here carbon content) of a spherical particle described as:

$$m(d) = \alpha d^3 \quad (3)$$

where $\alpha = \pi\rho/6$, ρ is the average density of the particle, and $w(d)$ is the settling rate calculated using Stokes Law:

$$w(d) = \beta d^2 \quad (4)$$

where $\beta = g(\rho - \rho_0)(18\nu\rho_0)^{-1}$, g is the gravitational acceleration, ρ_0 the fluid density, and ν the kinematic viscosity.

In addition, mass and settling rates of particles, $m(d)$ and $w(d)$, respectively, are often described as power law functions of their diameter obtained by fitting observed data, $m(d) \cdot w(d) = Ad^B$. The particles carbon flux can then be estimated using an approximation of equation (2) over a finite number (x) of small logarithmic intervals for diameter d spanning from 250 μm to 1.5 mm (particles <250 μm and >1.5 mm are not considered, consistent with the method presented in ref. 56) such as

$$F = \sum_{i=1}^x n_i A d_i^B \Delta d_i \quad (5)$$

where $A = 12.5 \pm 3.40$ and $B = 3.81 \pm 0.70$ have been estimated using a global data set that compared particle fluxes in sediment traps and particle size distributions from the UVP images.

Genomic data collection. For the sake of consistency between all available data sets from the *Tara* Oceans expeditions, we considered subsets of the data recently published in Science^{23–25}. In brief, one sample corresponds to data collected at one depth (surface (SRF) or deep chlorophyll maximum (DCM) determined from the profile of chlorophyll fluorometer) and at one station. To study the eukaryotic community in our current manuscript, we selected stations at which we had environmental data and carbon export estimated at 150 m with the UVP and all size fractions. Consequently a subset of 33 stations (corresponding to 56 samples) has been created compared to the 47 stations analysed in ref. 24. A similar procedure has been applied to the prokaryotic and viral data sets, reducing the prokaryotic

data set from ref. 23 to a subset of 104 samples from 62 stations and the viral data set from ref. 25 into a subset of 37 samples from 22 stations (See Supplementary Table 10). In addition a detailed table is provided summarizing which samples (depth and station) are available for each domain (Supplementary Table 11).

Eukaryotic taxa profiling. Photic-zone eukaryotic plankton diversity has been investigated through millions of environmental Illumina reads. Sequences of the 18S ribosomal RNA gene V9 region were obtained by PCR amplification and a stringent quality-check pipeline has been applied to remove potential chimaera or rare sequences (details on data cleaning in ref. 24). For 47 stations, and if possible at two depths (SRF and DCM), eukaryotic communities were sampled in the piconano- (0.8–5 μm), micro- (20–180 μm) and mesoplankton (180–2,000 μm) fractions (a detailed list of these samples is given in Supplementary Table 12). In the framework of the carbon export study, sequences from all size fractions were pooled in order to get the most accurate and statistically reliable data set of the eukaryotic community. The 2.3 million eukaryotic ribotypes were assigned to known eukaryotic taxonomic entities by global alignment to a curated database²⁴. To get the most accurate vision of the eukaryotic community, sequences showing less than 97% identity with reference sequences were excluded. The final eukaryotic relative abundance matrix used in our analyses included 1,750 lineages (taxonomic assignment has been performed using a last common ancestor methodology, and had thus been performed down to species level when possible) in 56 samples from 33 stations. Pooled abundance (number of V9 sequences) of each lineage has been normalized by the total sum of sequences in each sample.

Prokaryotic taxa profiling. To investigate the prokaryotic lineages, communities were sampled in the picoplankton. Both filter sizes have been used along the *Tara* Oceans transect: up to station #52, prokaryotic fractions correspond to a 0.22–1.6 μm size fraction, and from station #56, prokaryotic fractions correspond to a 0.22–3 μm size fraction. Prokaryotic taxonomic profiling was performed using 16S rRNA gene tags directly identified in Illumina-sequenced metagenomes (mitags) as described in ref. 57. 16S mitags were mapped to cluster centroids of taxonomically annotated 16S reference sequences from the SILVA database⁵⁸ (release 115: SSU Ref NR 99) that had been clustered at 97% sequence identity using USEARCH v. 6.0.307⁵⁹. 16S mitag counts were normalized by the total reads count in each sample (further details in ref. 23). The photic-zone prokaryotic relative abundance matrix used in our analyses included 3,253,962 mitags corresponding to 1,328 genera in 104 samples from 62 stations.

Prokaryotic functional profiling. For each prokaryotic sample, gene relative abundance profiles were generated by mapping reads to the OM-RGC using the MOCAT pipeline⁶⁰. The relative abundance of each reference gene was calculated as gene-length-normalized base counts. And functional abundances were calculated as the sum of the relative abundances of these reference genes, annotated to OG functional groups. In our analyses, we used the subset of the OM-RGC that was annotated to Bacteria or Archaea (24.4 million genes). Using a rarefied (to 33 million inserts) gene count table, an OG was considered to be part of the ocean microbial core if at least one insert from each sample was mapped to a gene annotated to that OG. For further details on the prokaryotic profiling please refer to ref. 23. The final prokaryotic functional relative abundance matrix used in our analyses included 37,832 OGs or functions in 104 samples from 62 stations. Genes from functions of FNET1 and FNET2 subnetworks were taxonomically annotated using a modified dual BLAST-based last common ancestor (2bLCA) approach⁶¹. We used RAPsearch2⁶² rather than BLAST to efficiently process the large data volume and a database of non-redundant protein sequences from UniProt (version: UniRef_2013_07) and eukaryotic transcriptome data not represented in UniRef (see Supplementary Tables 5 and 6, for full annotations).

Enumeration of prokaryotes by flow cytometry. For prokaryote enumeration by flow cytometry, three aliquots of 1 ml of seawater (pre-filtered by 200- μm mesh) were collected from both SRF and DCM. The samples were fixed immediately using cold 25% glutaraldehyde (final concentration 0.125%), left in the dark for 10 min at room temperature, flash-frozen and kept in liquid nitrogen on board and then stored at -80°C on land. Two subsamples were taken to separate counts of heterotrophic prokaryotes (not shown herein) and phototrophic picoplankton. For heterotrophic prokaryote determination, 400 μl of sample was added to a diluted SYTO-13 (Molecular Probes Inc.) stock (10:1) at $2.5 \mu\text{mol l}^{-1}$ final concentration, left for about 10 min in the dark to complete the staining and run in the flow cytometer. We used a FACS Calibur (Becton & Dickinson) flow cytometer equipped with a 15 mW argon-ion laser (488 nm emission). At least 30,000 events were acquired for each subsample (usually 100,000 events). Fluorescent beads (1 μm , Fluoresbrite carboxylate microspheres, Polysciences Inc.) were added at a known density as internal standards. The bead standard concentration was determined by epifluorescence microscopy. For phototrophic picoplankton, we used the same procedure as for heterotrophic prokaryote, but without addition of SYTO-13. Data analysis was performed with FlowJo software (Tree Star, Inc.).

Profiling of viral populations. In order to associate viruses to carbon export we used viral populations as defined in ref. 25 using a set of 43 *Tara* Oceans viromes. In brief, viral populations were defined as large contigs (>10 predicted genes and >10 kb) identified as most likely originating from bacterial or archaeal viruses. These 6,322 contigs remained and were then clustered into populations if they shared more than 80% of their genes at >95% nucleotide identity. This resulted in 5,477 'populations' from the 6,322 contigs, where as many as 12 contigs were included per population. For each population, the longest contig was chosen as the 'seed' representative sequence. The relative abundance of each population was computed by mapping all quality-controlled reads to the set of 5,477 non-redundant populations (considering only mapping quality scores greater than 1) with Bowtie2 (ref. 63) and if more than 75% of the reference sequence was covered by virome reads. The relative abundance of a population in a sample was computed as the number of base pairs recruited to the contig normalized to the total number of base pairs available in the virome and the contig length if more than 75% of the reference sequence was covered by virome reads, and set to 0 otherwise (see ref. 25 for further details). The final viral population abundance matrix used in our analyses included 5,291 viral population contigs in 37 samples from 22 stations.

Viral host predictions. The longest contig in a population was defined as the seed sequence and considered the best estimate of that population's origin. These seed sequences were used to assess taxonomic affiliation of each viral population. Cases where >50% of the genes were affiliated to a specific reference genome from RefSeq Virus (based on a BLASTP comparison with thresholds of 50 for bit score and 1×10^{-5} for e-value) with an identity percentage of at least 75% (at the protein sequence level) were considered as confident affiliations to the corresponding reference virus. The viral population host group was then estimated based on these confident affiliations (see Supplementary Table 13 for host affiliation of viral population contigs associated to carbon export).

Viral protein clusters. Viral protein clusters (PCs) correspond to ORFs initially mapped to existing clusters (POV, GOS and phage genomes). The remaining, unmapped ORFs were self-clustered, using cd-hit as described in ref. 25. Only PCs with more than two ORFs were considered bona fide and were used for subsequent analyses. To compute PC relative abundance for statistical analyses, reads were mapped back to predicted ORFs in the contigs data set using Mosaik as described in ref. 25. Read counts to PCs were normalized by sequencing depth of each virome. Importantly, we restricted our analyses to 4,294 PCs associated to the 277 viral population contigs significantly associated to carbon export in 37 samples from 22 stations.

Sparse partial least squares analysis. In order to directly associate eukaryotic lineages to carbon export and other environmental traits (Fig. 1b), we used sparse partial least square (sPLS)⁶⁴ as implemented in the R package mixOmics²⁹. We applied the sPLS in regression mode, which will model a causal relationship between the lineages and the environmental traits, that is, PLS will predict environmental traits (for example, carbon export) from lineage abundances. This approach enabled us to identify high correlations (see Supplementary Table 1) between certain lineages and carbon export but without taking into account the global structure of the planktonic community.

Co-occurrence network model analysis. Weighted correlation network analysis (WGCNA) was performed to delineate feature (lineages, viral populations, PCs or functions) subnetworks based on their relative abundance^{65,66}. A signed adjacency measure for each pair of features was calculated by raising the absolute value of their Pearson correlation coefficient to the power of a parameter p . The default value $p = 6$ was used for each global network, except for the Prokaryotic functional network where p had to be lowered to 4 in order to optimize the scale-free topology network fit. Indeed, this power allows the weighted correlation network to show a scale-free topology where key nodes are highly connected with others. The obtained adjacency matrix was then used to calculate the topological overlap measure (TOM), which for each pair of features, taking into account their weighted pairwise correlation (direct relationships) and their weighted correlations with other features in the network (indirect relationships). For identifying subnetworks a hierarchical clustering was performed using a distance based on the TOM measure. This resulted in the definition of several subnetworks, each represented by its first principal component.

These characteristic components play a key role in weighted correlation network analysis. On the one hand, the closeness of each feature to its cluster, referred to as the subnetwork membership, is measured by correlating its relative abundance with the first principal component of the subnetwork. On the other hand, association between the subnetworks and a given trait is measured by the pairwise Pearson correlation coefficients between the considered environmental trait and their respective principal components. A similar protocol has been performed on the eukaryotic relative abundance matrix, the prokaryotic relative abundance matrix, the prokaryotic functions relative abundance matrix and the viral

population and PC relative abundance matrices. All procedures were applied on Hellinger-transformed log-scaled abundances. Notably, the protocol is not sensitive to copy number variation as observed across different eukaryotic species, because the association between two species relies on a correlation score between relative abundance measurements. Computations were carried out using the R package WGCNA³³.

Given the nature of the eukaryotic data set (three distinct size fractions), the sampling process may lead to the loss of size fractions. In particular, samples 1, 3, 17, 37, 39, 43, 48, 53, 54, 55 and 66 are eventually biased by such a loss (Supplementary Table 12). A complementary WGCNA analysis was performed with addition of these samples to evaluate the robustness of our protocol to missing size fractions. The composition of the eukaryotic subnetwork built with an extended data set (that is, 67 samples from 37 stations for which size fractions were missing in 11 samples) was compared to the subnetwork as presented above (that is, 56 samples from 33 stations). Both subnetworks show an overlap of 75% of lineage, whereas four of the top five VIP lineages with the extended data set (see Extended Data Fig. 5 for details) can be found in the top six VIP lineages of the above subnetwork (Supplementary Table 2), emphasizing highly similar results and a small sensitivity to size fraction loss.

Extraction of subnetworks related to carbon export. For each subnetwork (called modules within WGCNA) extracted from each global network, pairwise Pearson correlation coefficients between the subnetwork principal components and the carbon export estimation was computed, as well as corresponding P values corrected for multiple testing using the Benjamini and Hochberg FDR procedure. The subnetworks showing the highest correlation scores are of interest and were investigated. One subnetwork (49 nodes) was significant within the eukaryotic network; one subnetwork (109 nodes) was significant for the prokaryotic network; one subnetwork (277 nodes) was significant within the virus network; two subnetworks (441 and 220 nodes) were significant within the prokaryotic functional network, and two subnetworks (1,879 and 2,147 nodes) were significant within the viral PCs network.

Partial least squares regression. In addition to the network analyses, we asked whether the identified subnetworks can be used as predictors for the carbon export estimations. To answer this question, we used partial least squares (PLS) regression, which is a dimensionality-reduction method that aims at determining predictor combinations with maximum covariance with the response variable. The identified combinations, called latent variables, are used to predict the response variable. The predictive power of the model is assessed by correlating the predicted vector with the measured values. The significance of the prediction power was evaluated by permuting the data 10,000 times. For each permutation, a PLS model was built to predict the randomized response variable and a Pearson correlation was calculated between the permuted response variable and in leave-one-out cross-validation (LOOCV) predicted values. The 10,000 random correlations are compared to the performance of the PLS model that were used to predict the true response variable. In addition, the predictors were ranked according to their value importance in projection (VIP)⁶⁷. The VIP measure of a predictor estimates its contribution in the PLS regression. The predictors having high VIP values are assumed important for the PLS prediction of the response variable. The VIP values of the prokaryotic functional subnetworks are provided in Supplementary Tables 5, 6. For the sake of illustration, only lineages or functions with VIP > 1 (ref. 67) are discussed and pictured in Figs 2 and 4. Our computations were carried out using the R package pls⁶⁸. All programs are available under GPL Licence.

Subnetwork representations. Nodes of the subnetworks represent either lineages (eukaryotic, prokaryotic or viral) or functions (prokaryotic or viral). Subnetworks related to the carbon export have been represented in two distinct formats. Scatter plots represent each nodes based on their Pearson correlation to the carbon export and their respective node centrality within the subnetwork. The latter has been recomputed using significant Spearman correlations above 0.3 (>0.9 for viral PCs) as edges, this is done for visualization purposes since WGCNA subnetworks (based on the topology overlap measure (TOM) between nodes) are hyper-connected. Size representation of nodes are proportional to the VIP score after PLS. The hive plots depict the same subnetworks by focusing on two main features: x axis and y axis depict nodes of subnetworks ranked by their VIP scores and Pearson correlation to the carbon export, respectively.

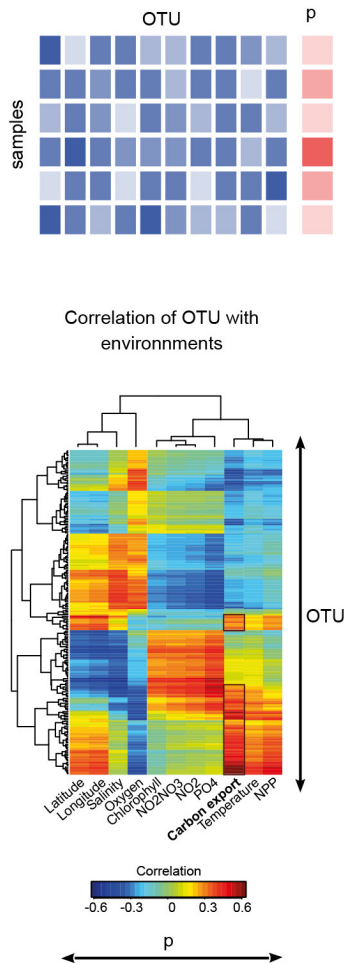
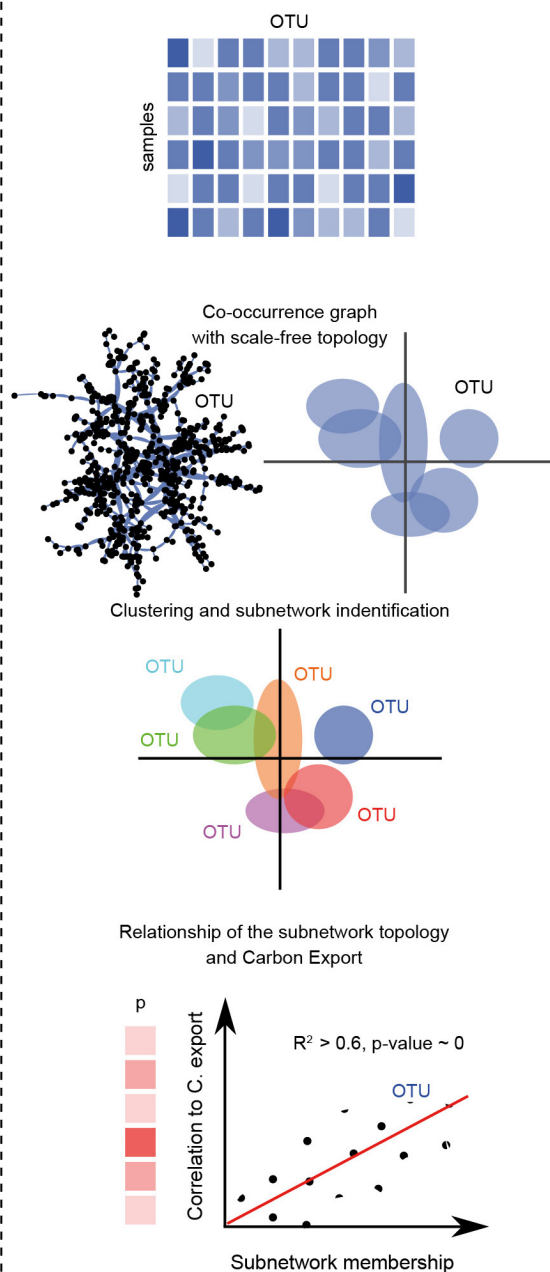
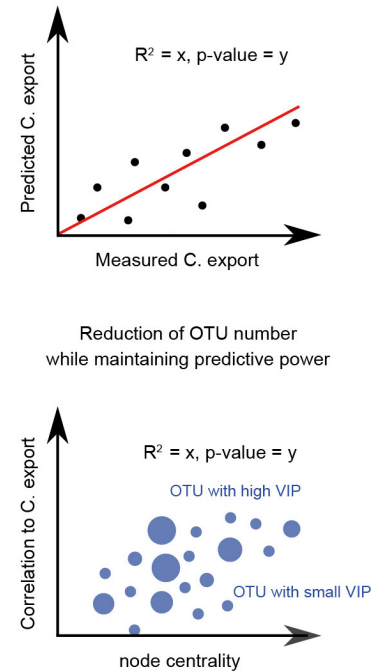
52. Picheral, M. et al. The Underwater Vision Profiler 5: An advanced instrument for high spatial resolution studies of particle size spectra and zooplankton. *Limnol. Oceanogr. Methods* **8**, 462–473 (2010).

53. Behrenfeld, M. J. & Falkowski, P. G. Photosynthetic rates derived from satellite-based chlorophyll concentration. *Limnol. Oceanogr.* **42**, 1–20 (1997).

54. McCave, I. N. Size spectra and aggregation of suspended particles in the deep ocean. *Deep Sea Res. Part I Oceanogr. Res. Pap.* **31**, 329–352 (1984).

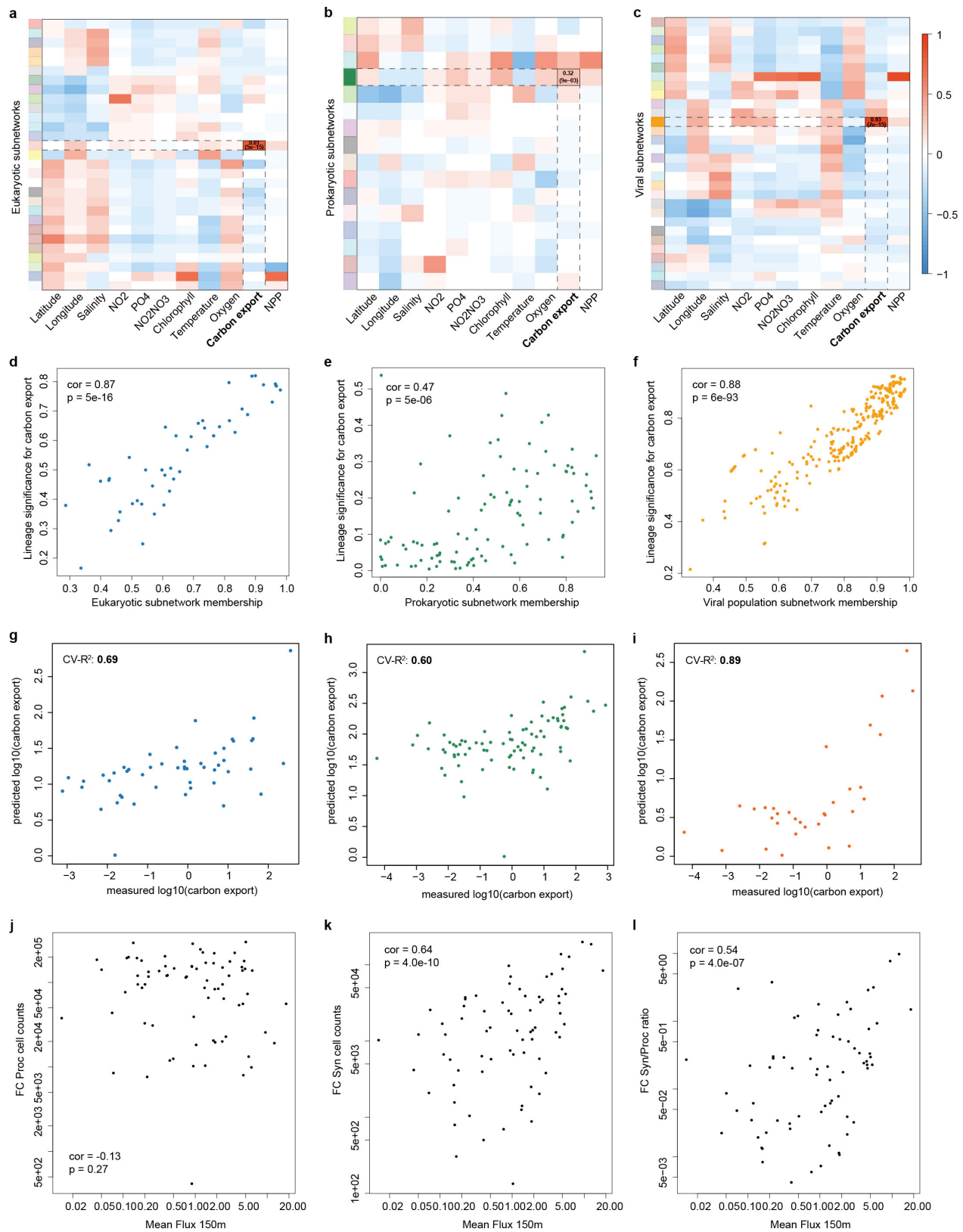
55. Sheldon, R. W., Prakash, A. & Sutcliffe, W. H. Size distribution of particles in ocean. *Limnol. Oceanogr.* **17**, 327–340 (1972).

56. Guidi, L. *et al.* Relationship between particle size distribution and flux in the mesopelagic zone. *Deep Sea Res. Part I Oceanogr. Res. Pap.* **55**, 1364–1374 (2008).
57. Logares, R. *et al.* Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ. Microbiol.* **16**, 2659–2671 (2014).
58. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
59. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
60. Kultima, J. R. *et al.* MOCAT: a metagenomics assembly and gene prediction toolkit. *PLoS ONE* **7**, e47656 (2012).
61. Hingamp, P. *et al.* Exploring nucleo-cytoplasmic large DNA viruses in *Tara* Oceans microbial metagenomes. *ISME J.* **7**, 1678–1695 (2013).
62. Zhao, Y., Tang, H. & Ye, Y. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* **28**, 125–126 (2012).
63. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).
64. Shen, H. P. & Huang, J. H. Z. Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivar. Anal.* **99**, 1015–1034 (2008).
65. Langfelder, P. & Horvath, S. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst. Biol.* **1**, 54 (2007).
66. Li, A. & Horvath, S. Network neighborhood analysis with the multi-node topological overlap measure. *Bioinformatics* **23**, 222–231 (2007).
67. Chong, I. G. & Jun, C. H. Performance of some variable selection methods when multicollinearity is present. *Chemometr. Intell. Lab.* **78**, 103–112 (2005).
68. Mevik, B. H. & Wehrens, R. The pls package: Principal component and partial least squares regression in R. *J. Stat. Softw.* **18**, 1–23 (2007).

a Pairwise approach**b** Graph-based approach (WGCNA)**c** Machine learning technique (PLS)**Extended Data Figure 1 | Overview of analytical methods used**

in the manuscript. **a**, Depiction of a standard pairwise analysis that considers a sequence relative abundance matrix for s samples ($s \times \text{OTUs}$ (operational taxonomic units)) and its corresponding environmental matrix ($s \times p$ (parameters)). sPLS results emphasize OTU(s) that are the most correlated to environmental parameters. **b**, Depiction of a graph-based approach. Using only a relative abundance matrix ($s \times \text{OTUs}$), WGCNA builds a graph where nodes are OTUs and edges represent significant co-occurrence. Co-occurrence scores between nodes are weights allocated to corresponding edges. These weights are magnified by a power-law function until the graph becomes scale-free. The graph is then decomposed within subnetworks (groups of OTUs) that are analysed separately. One subnetwork (group of OTUs) is considered of interest when its topology is related to the trait of interest; in the current case

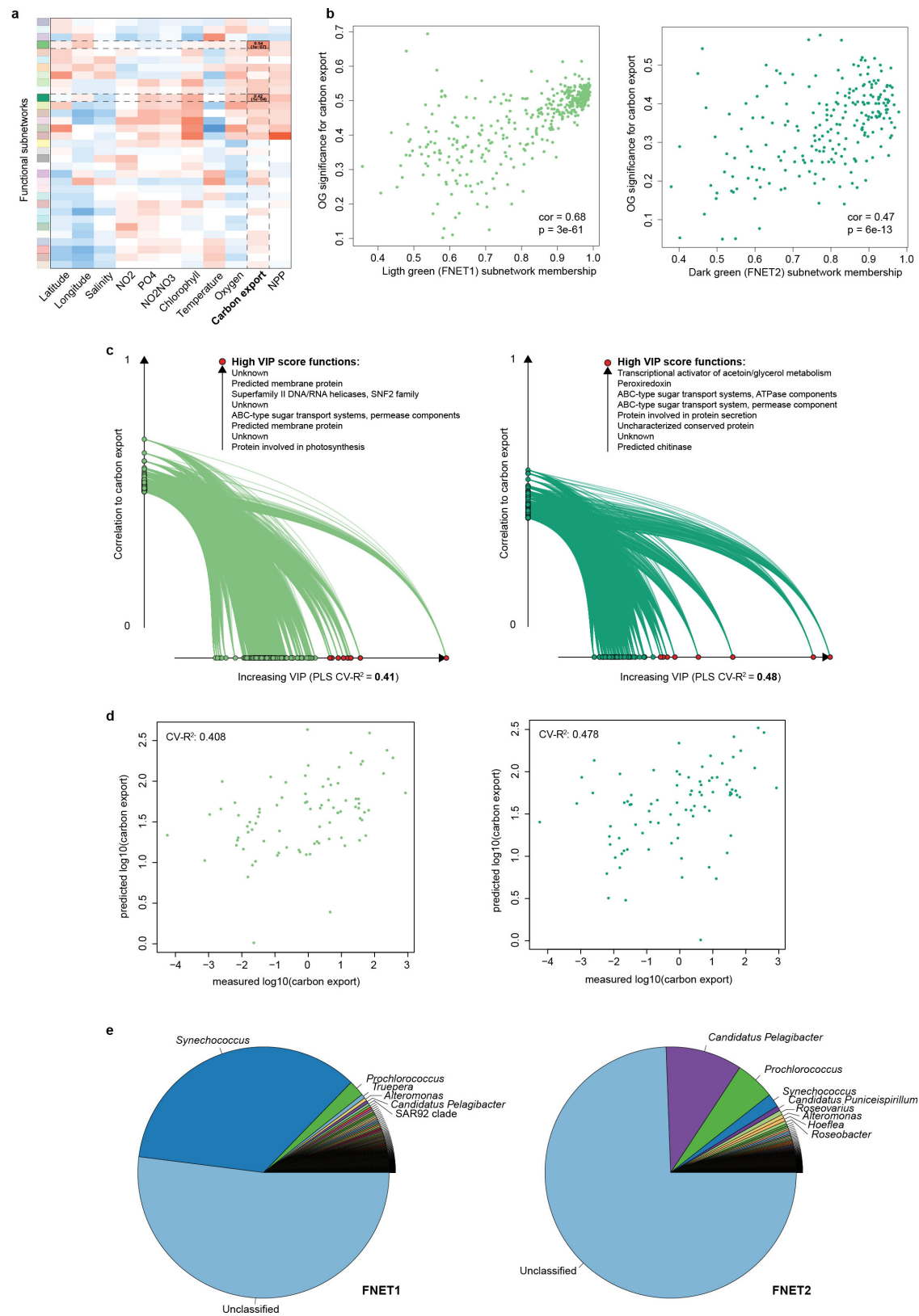
carbon export. For each subnetwork (for instance the subnetwork related to carbon export), each OTU is spread within a feature space that plots each OTU based on its membership to the subnetwork (x axis) and its correlation to the environmental trait of interest (that is, carbon export). A good regression of all OTUs emphasizes the putative relation of the subnetwork topology and the carbon export trait (that is, the more a given OTU defines the subnetwork topology, the more it is correlated to carbon export). **c**, Depiction of the machine learning (PLS) approach that was applied following subnetwork identification and selection. Greater VIP scores (that is, larger circles) emphasized most important OTUs. VIP refers to variable importance in projection and reflects the relative predictive power of a given OTU. OTUs with a VIP score greater than 1 are considered as important in the predictive model and their selection does not alter the overall predictive power.



Extended Data Figure 2 | See next page for figure caption.

Extended Data Figure 2 | Lineage ecological subnetworks associated to environmental parameters and their structures correlating to carbon export. **a–c**, Global ecological networks were built using the WGCNA methodology (see Methods) and correlated to classical oceanographic parameters as well as carbon export (estimated at 150 m from particle size distribution and abundance). Each domain-specific global network is decomposed into smaller coherent subnetworks (depicted by distinct colours on the y axis) and their eigenvector is correlated to all environmental parameters. Similar to a correlation at the network scale, this approach directly links subnetworks to environmental parameters (that is, the more the taxa contribute to the subnetwork structure, the more their abundance is correlated to the parameter). **a**, A single eukaryotic subnetwork ($n = 58$, $N = 1,870$) is strongly associated to carbon export ($r = 0.81$, $P = 5 \times 10^{-15}$). **b**, A single prokaryotic subnetwork ($n = 109$, $N = 1,527$) is moderately associated to carbon export ($r = 0.32$, $P = 9 \times 10^{-3}$). **c**, A single viral subnetwork ($n = 277$, $N = 5,476$) is strongly associated to carbon export ($r = 0.93$, $P = 2 \times 10^{-15}$). **d–f**, The WGCNA approach directly links subnetworks to environmental parameters, that is, the more the features contribute to the subnetwork structure (topology), the more their abundance are correlated to the parameter.

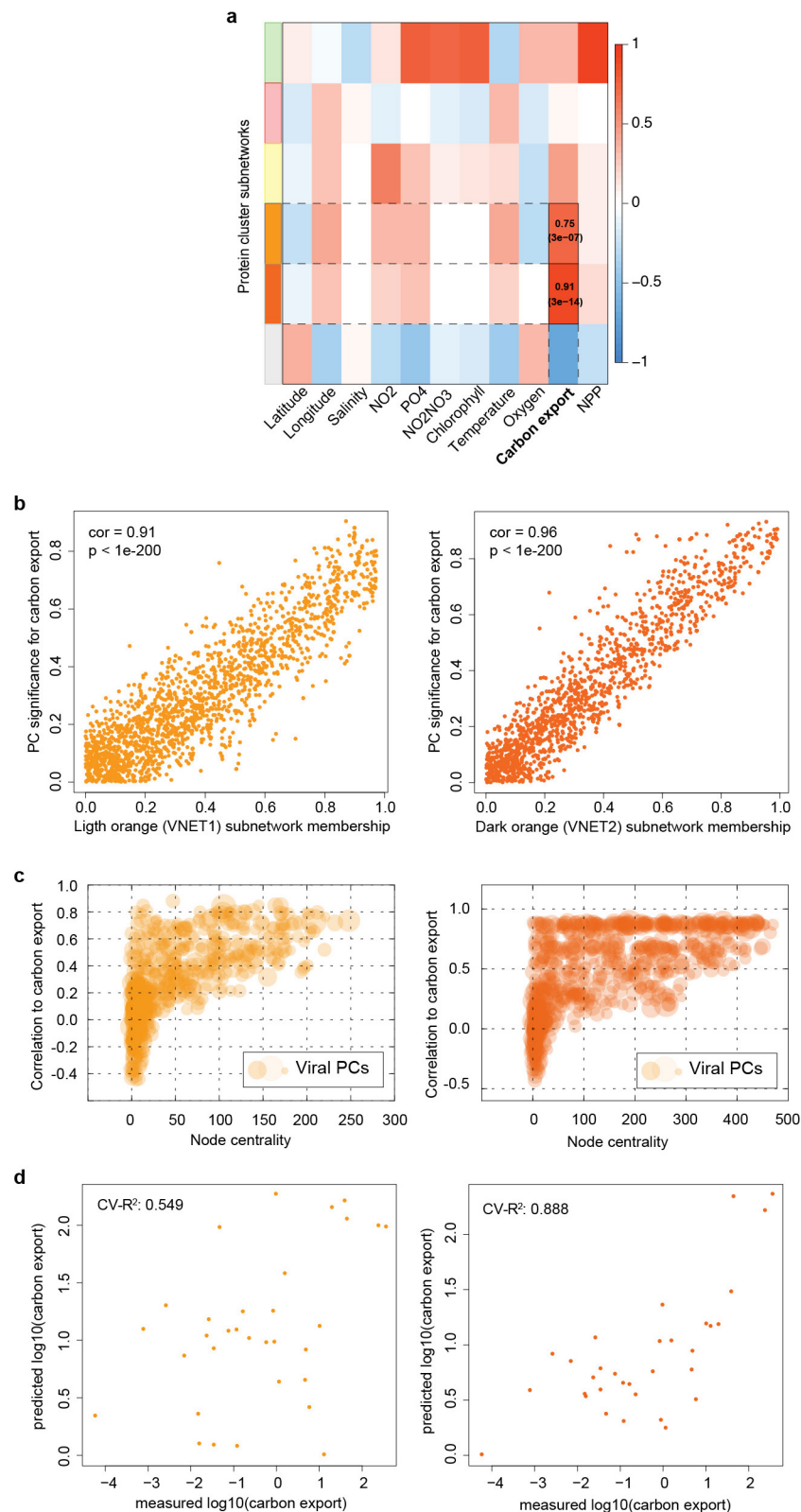
This measure allows to identify subnetworks for which the overall structure, summarized as the eigenvector of the subnetwork, is related to the carbon export. **d**, The eukaryotic subnetwork structure correlates to carbon export ($r = 0.87$, $P = 5 \times 10^{-16}$). **e**, The prokaryotic subnetwork structure correlates to carbon export ($r = 0.47$, $P = 5 \times 10^{-6}$). **f**, The viral population subnetwork structure correlates to carbon export ($r = 0.88$, $P = 6 \times 10^{-93}$). **g–i**, Lineage subnetworks predict carbon export. PLS regression was used to predict carbon export using lineage abundances in selected subnetworks. LOOCV was performed and VIP scores computed for each lineage. **g**, The eukaryotic subnetwork predicts carbon export with a R^2 of 0.69. **h**, The prokaryotic subnetwork predicts carbon export with a R^2 of 0.60. **i**, The viral population subnetwork predicts carbon export with a R^2 of 0.89. **j–l**, *Synechococcus* (rather than *Prochlorococcus*) absolute cell counts correlate well to carbon export. **j**, *Prochlorococcus* cell counts estimated by flow cytometry do not correlate to carbon export (mean carbon flux at 150 m, $r = -0.13$, $P = 0.27$). **k**, *Synechococcus* cell counts estimated by flow cytometry correlate significantly to carbon export ($r = 0.64$, $P = 4.0 \times 10^{-10}$). **l**, *Synechococcus* / *Prochlorococcus* cell counts ratio correlates significantly to carbon export ($r = 0.54$, $P = 4.0 \times 10^{-7}$).



Extended Data Figure 3 | See next page for figure caption.

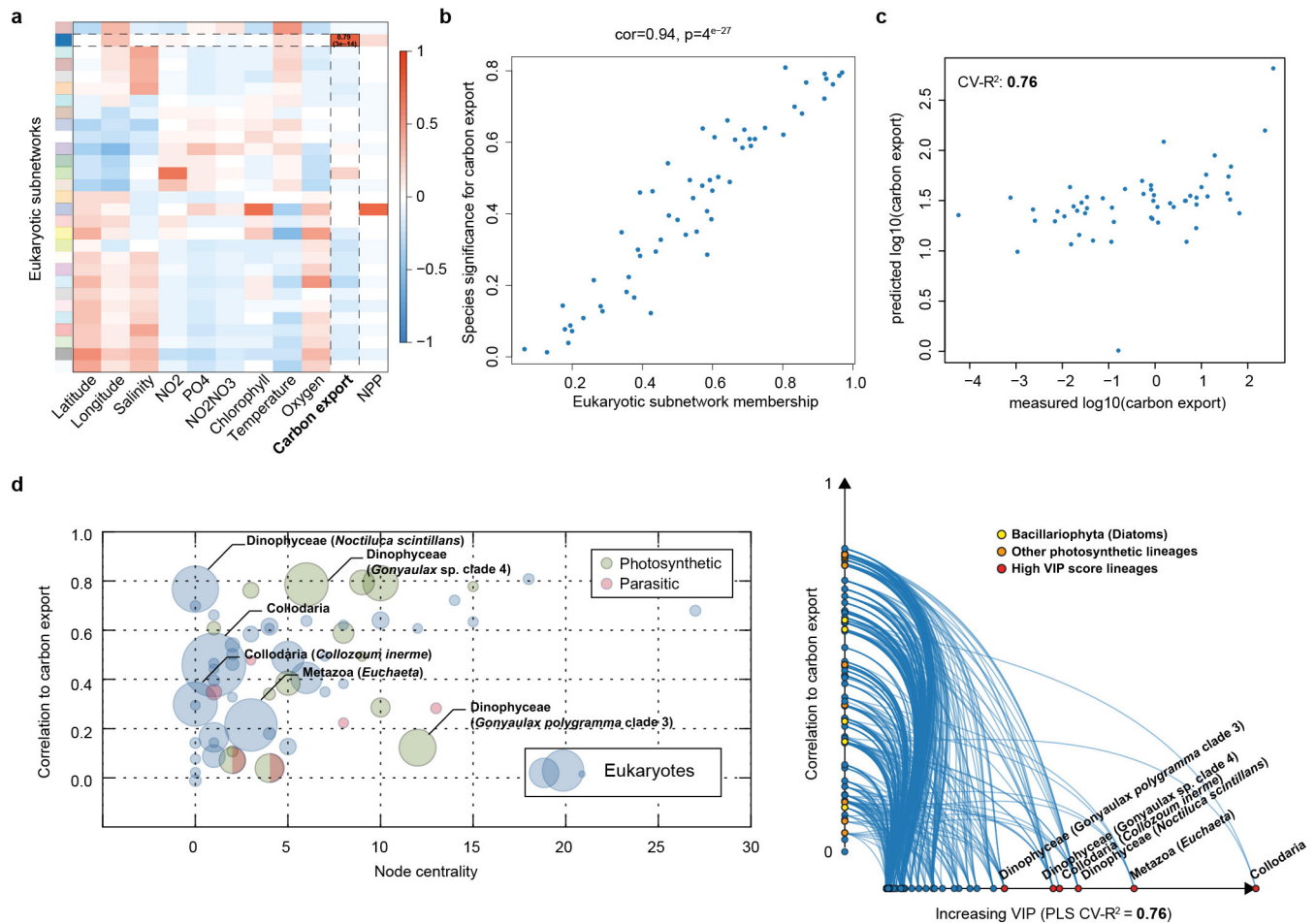
Extended Data Figure 3 | Prokaryotic function subnetworks associated to environmental parameters and their structure correlate to carbon export. **a–c**, Global ecological networks were built for the prokaryotic functions using the WGCNA methodology (see Methods) and correlated to classical oceanographic parameters as well as carbon export. **a**, Two bacterial functional subnetworks ($n = 441$ and $n = 220$, $N = 37,832$) are associated to carbon export ($r = 0.54$, $P = 1 \times 10^{-7}$ and $r = 0.42$, $P = 1 \times 10^{-4}$). **b**, The WGCNA approach directly links subnetworks to environmental parameters, that is, the more the features contribute to the subnetwork structure (topology), the more their abundance are correlated to the parameter. This measure allows to identify subnetworks for which the overall structure, summarized as the eigenvector of the subnetwork, is related to the carbon export. The bacterial function subnetwork structures correlate to carbon export (FNET1 $r = 0.68$, $P = 3 \times 10^{-61}$, and FNET2 $r = 0.47$, $P = 6 \times 10^{-13}$). **c**, Two functional subnetworks (light and dark green, FNET1 ($n = 220$) and FNET2 ($n = 441$), respectively)

are significantly associated with carbon export (FNET1: $r = 0.42$, $P = 4 \times 10^{-9}$ and FNET2: $r = 0.54$, $P = 7 \times 10^{-6}$). The highest VIP score functions from top to bottom correspond to red dots from right to left. **d**, PLS regression was used to predict carbon export using abundances of functions (OGs) in selected subnetworks. LOOCV was performed and VIP scores computed for each function. Light green subnetwork (FNET1) functions predict carbon export with a R^2 of 0.41. Dark green subnetwork (FNET2) functions predict carbon export with a R^2 of 0.48. **e**, Cumulative abundance of genus-level taxonomic annotations of genes encoding functions from FNET1 and FNET2 subnetworks and bacterial function subnetworks predict carbon export. Genes contributing to the relative abundance of FNET1 and FNET2 subnetwork functions were taxonomically annotated by homology searches against a non-redundant gene reference database using a last common ancestor (LCA) approach (see Methods).



Extended Data Figure 4 | Viral protein cluster networks reveal potential marker genes for carbon export prediction at global scale. **a**, A viral protein cluster (PC) network was built using abundances of PCs predicted from viral population contigs associated to carbon export (Fig. 2c) using the WGCNA methodology (see Methods) and correlated to classical oceanographic parameters. Two viral PC subnetworks ($n = 1,879$ and $n = 2,147$, $N = 4,678$, light and dark orange, VNET1 and VNET2, left and right panel respectively) are strongly associated to carbon export (VNET1: $r = 0.75$, $P = 3 \times 10^{-7}$ and VNET2: $r = 0.91$, $P = 3 \times 10^{-14}$). **b**, The viral

PC subnetwork structures correlate to carbon export (VNET1 $r = 0.91$, $P < 1 \times 10^{-200}$, and VNET2 $r = 0.96$, $P < 1 \times 10^{-200}$). **c**, Size of dots is proportional to the VIP score computed for the PLS regression. **d**, Viral PC subnetworks predict carbon export. PLS regression was used to predict carbon export using abundances of viral protein clusters (PCs) in selected subnetworks. LOOCV was performed and VIP scores computed for each PC. Light orange subnetwork (VNET1, left panel) PCs predict carbon export with a R^2 of 0.55. Dark orange subnetwork (VNET2, right panel) PCs predict carbon export with a R^2 of 0.89.



Extended Data Figure 5 | WGCNA and PLS regression analyses for the full eukaryotic data set. **a**, A single eukaryotic subnetwork ($n=58$), is strongly associated to carbon export ($r=0.79$, $P=3 \times 10^{-14}$). **b**, The eukaryotic subnetwork structure correlates to carbon export ($r=0.94$, $P=4 \times 10^{-27}$). **c**, The eukaryotic subnetwork predicts carbon export with a

R^2 of 0.76. **d**, Lineages with the highest VIP score (dot size is proportional to the VIP score in the scatter plot) in the PLS are depicted as red dots corresponding to two rhizaria (Collocladia), one copepod (*Euchaeta*), and three dinophyceae (*Noctiluca scintillans*, *Gonyaulax polygramma* and *Gonyaulax* sp. (clade 4)).

Dynamics from noisy data with extreme timing uncertainty

R. Fung^{1*}, A. M. Hanna^{2,3,4}, O. Vendrell^{2,3}, S. Ramakrishna⁵, T. Seideman⁵, R. Santra^{2,3,4,6} & A. Ourmazd^{1*}

Imperfect knowledge of the times at which ‘snapshots’ of a system are recorded degrades our ability to recover dynamical information, and can scramble the sequence of events. In X-ray free-electron lasers, for example, the uncertainty—the so-called timing jitter—between the arrival of an optical trigger (‘pump’) pulse and a probing X-ray pulse can exceed the length of the X-ray pulse by up to two orders of magnitude¹, marring the otherwise precise time-resolution capabilities of this class of instruments. The widespread notion that little dynamical information is available on timescales shorter than the timing uncertainty has led to various hardware schemes to reduce timing uncertainty^{2–4}. These schemes are expensive, tend to be specific to one experimental approach and cannot be used when the record was created under ill-defined or uncontrolled conditions such as during geological events. Here we present a data-analytical approach, based on singular-value decomposition and nonlinear Laplacian spectral analysis^{5–7}, that can recover the history and dynamics of a system from a dense collection of noisy snapshots spanning a sufficiently large multiple of the timing uncertainty. The power of the algorithm is demonstrated by extracting the underlying dynamics on the few-femtosecond timescale from noisy experimental X-ray free-electron laser data recorded with 300-femtosecond timing uncertainty¹. Using a noisy dataset from a pump-probe experiment on the Coulomb explosion of nitrogen molecules, our analysis reveals vibrational wave-packets consisting of components with periods as short as 15 femtoseconds, as well as more rapid changes, which have yet to be fully explored. Our approach can potentially be applied whenever dynamical or historical information is tainted by timing uncertainty.

The fundamental premise of our approach is simple. A series of snapshots concatenated in the order of their inaccurate time stamps will contain some time-evolutionary information (‘a weak arrow of time’), provided that the concatenation window spans a period comparable with, or longer than, the timing uncertainty associated with each individual snapshot. This realization leads one to consider a series of c -fold concatenated snapshots, formed by moving a c -frame-wide window over the raw dataset ordered according to the inaccurate time stamps. The dynamical history can then be extracted from the series of concatenated snapshots using techniques developed to extract signal from noise, such as singular-value decomposition (SVD)⁸. SVD determines a series of statistically significant modes, each consisting of a characteristic pattern (topogram) and its time evolution (chronogram). A topogram can be a characteristic image or spectrum, with the corresponding chronogram showing its change with time. For each mode, a singular value specifies the power contained in that mode⁸.

Consider snapshots, such as images or spectra, that can be represented as vectors by using the pixel values of each snapshot as the components of a vector \mathbf{x} . A snapshot can then be thought of as a point in multidimensional space, and a dataset as a cloud of points in that space.

Similar to principal component analysis, SVD is a linear-algebraic approach, efficiently applicable only when the data cloud defines a flat, low-dimensional hypersurface. Unfortunately, many systems of interest cannot be adequately treated within the framework of linear-algebraic methods such as SVD. Geometrically, data from such systems give rise to intrinsically curved hypersurfaces (manifolds). Fundamental to our approach, therefore, is nonlinear Laplacian spectral analysis (NLSA)⁶, which performs the same analysis as SVD, but on curved manifolds.

For a dataset consisting of a series of N_s time-ordered snapshots, the analysis begins with a ‘time-lagged embedding’^{9–11} to form c -fold concatenated ‘superframes’ (or ‘supervectors’) from the dataset consisting of vectors \mathbf{x} . A typical supervector

$$\mathbf{X}_i = (\mathbf{x}_i; \mathbf{x}_{i-\delta t}; \dots; \mathbf{x}_{i-(c-1)\delta t})$$

is obtained by appending the column vectors $\mathbf{x}_{i-i\delta t}$ ($0 \leq i \leq c-1$) to each other, with $\mathbf{x}_{i-i\delta t}$ representing the i th in the sequence of c snapshots ordered according to time stamps. The time stamp assigned to each of the resulting $(N_s - c)$ supervectors is defined as the mean of the time stamps of its constituent vectors. Unlike averaging, concatenation retains the information content of the dataset⁶; see Supplementary Information section 1.

Next, we use graph-based analysis—specifically the diffusion map algorithm⁵—to identify the nonlinear data manifold formed by the collection of supervectors. The matrix X of supervectors \mathbf{X}_i is then projected (in the sense defined in ref. 6) onto the manifold, to obtain the matrix A

$$A = X\mu\Phi \quad (1)$$

with μ the Riemannian measure of the manifold and Φ the empirical orthogonal eigenfunctions (EOFs)—a truncated set of the eigenfunctions of the Laplace–Beltrami operator on the manifold⁶. This Euclidean description of the nonlinear manifold allows us to analyse the matrix A using standard SVD. The chronograms obtained using SVD are projected from the space defined by Φ back to the time domain, and the topograms corresponding to the superframes are ‘unwrapped’ to obtain individual frames⁶. This approach is able to deal naturally with complex nonlinear dynamics^{6,12}, and to extract conformational information from ultralow-signal snapshots of molecular machines¹³.

Now consider the effect of stochastic timing uncertainty. Recall that the data matrix X is affected by timing uncertainty in two ways: first, the sequence of superframes can differ from the correct, jitter-free case; and, second, the time intervals within the members of a superframe, and those between the superframes themselves, can vary stochastically about a mean. It can be shown that the SVD step in NLSA is immune to jitter-induced changes in the superframe sequence, which are in any case unlikely for large concatenation parameters. As for non-uniformity

¹Department of Physics, University of Wisconsin Milwaukee, 3135 North Maryland Avenue, Milwaukee, Wisconsin 53211, USA. ²Center for Free-Electron Laser Science, DESY, Notkestrasse 85, 22607 Hamburg, Germany. ³The Hamburg Centre for Ultrafast Imaging, Luruper Chaussee 149, 22761 Hamburg, Germany. ⁴Department of Chemistry, University of Hamburg, Grindelallee 117, 20146 Hamburg, Germany. ⁵Department of Chemistry, Northwestern University, 2145 Sheridan Road, Evanston, Illinois 60208, USA. ⁶Department of Physics, University of Hamburg, Jungiusstrasse 9, 20355 Hamburg, Germany.

*These authors contributed equally to this work.

in time sampling, it can be shown analytically and by simulation that, with a sufficient number of snapshots, the outcome of SVD corresponds to time samples that are uniformly spaced to within small oscillations about the mean; see Supplementary Information section 2.

The results of SVD analysis of the matrix A must be projected back into the time domain to reconstruct the dynamics, which involves ‘undoing’ the effect of the projection described in equation (1) by computing $A\Phi^T$ (ref. 6). Because the EOFs represented by Φ are evaluated with our imperfect knowledge of timing, this back-projection re-injects jitter into the results. However, in the limit of large concatenation parameters, the manifold geometry and, hence, the EOFs are biased towards the most stable component of the dynamics⁷, as supported by the reduction in the number of eigenvalues above a spectral gap from five (in the manifold of raw data) to one (after concatenation); see Supplementary Information section 13. One may therefore regard the timing jitter as a form of stochastic forcing, which has been extensively studied¹⁴. In Supplementary Information section 3, we describe how reliable dynamical information can be obtained on timescales substantially shorter than the timing jitter.

For the experimental case analysed below (for which the full-width at half-maximum (FWHM) of the jitter is 300 fs; that is, $\sigma = 120$ fs, $c = 5,800$ and the average time-sample spacing is 50 as), the discussion in Supplementary Information section 3 leads us to expect reliable information on the femtosecond scale. This estimate ignores several important issues, such as the width of the probe pulse and the requirements of Shannon sampling; see Supplementary Information section 3. But it indicates that the approach we have outlined has the potential to reveal dynamics on timescales substantially shorter than the timing uncertainty. The effectiveness of this approach and the guidelines for

its use are outlined in Supplementary Information sections 4–7, with reference to four trial models.

We next demonstrate the capability to obtain dynamical information on timescales much shorter than the timing uncertainty using the noisy experimental data referred to above, which were recorded with substantial jitter stemming from the stochastic nature of the process used to generate ultrashort X-ray pulses in X-ray free-electron lasers¹⁵. The dataset, consisting of 10^5 time-of-flight spectral snapshots spanning a delay time from about -2.7 ps to $+2.3$ ps, was collected in the course of a pump-probe experiment on the Coulomb explosion of nitrogen molecules; see Supplementary Information section 12. During the experiment, an infrared pulse of approximately 60 fs in length either preceded or succeeded an ultrashort (< 10 fs) X-ray pulse generated by the X-ray free-electron laser at the Linac Coherent Light Source¹. Each time-of-flight spectrum (with a signal-to-noise ratio of approximately 0.16; see Supplementary Information section 10) is a record of the dynamics of about 30 molecular ions. The experimental approach and conditions are described in ref. 1.

The data from this experiment span three different regimes: first, that in which the X-ray pulse preceded the infrared pulse (the ‘X-ray-first’ regime); second, that in which the X-ray pulse succeeded the infrared pulse (the ‘infrared-first’ regime); and, third, that in which the ultrashort X-ray pulse arrives while the infrared pulse is active and so the two overlap (the ‘overlap’ regime). In the X-ray-first regime, the X-ray pulse ionizes (and/or dissociates) the N_2 molecules, and the infrared pulse breaks up the quasi-bound molecular ions it encounters, which are then collected by the time-of-flight spectrometer¹. In the infrared-first regime, the infrared pulse induces impulsive alignment¹⁶, and can also ionize the molecules, which are then dissociated by the X-ray

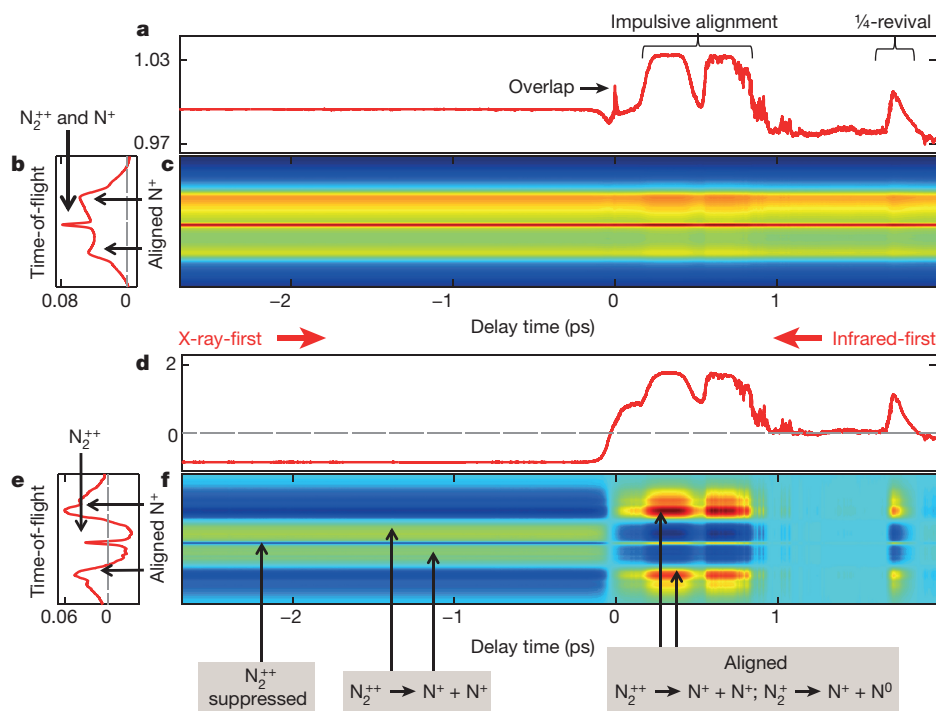


Figure 1 | Coulomb explosion of N_2 , singular modes 1 and 2. **a**, Mode-1 chronogram showing the time evolution of the average spectrum. The vertical axis shows the number of molecular fragments reaching the detector in arbitrary units. Note the three different regimes: X-ray-first, overlap and infrared-first. The effects of impulsive orientational alignment induced by the infrared pulse (including the 1/4-revival) are indicated. A sharp feature marks the overlap region, where the X-ray and infrared pulses overlap. **b**, Mode-1 time-of-flight spectrum showing the region around a mass-to-charge ratio of 14 (the horizontal axis shows the signal strength in arbitrary units). The sharp central peak is due to N_2^{++} dications. The side-peaks stem from N^+ ions ejected towards and away

from the detector. **c**, Mode-1 reconstructed time-of-flight series of spectral frames showing the evolution of the average spectrum with pump-probe delay time (the colour scale indicates the signal strength: lowest signal, dark blue; highest signal, red). **d**, Mode-2 chronogram showing the time evolution of the mode; vertical axis as in **a**. An error-function-like feature spans the region in which the X-ray and infrared pulses overlap. **e**, Mode-2 time-of-flight spectrum showing the region around a mass-to-charge ratio of 14 (horizontal axis as in **b**). **f**, Mode-2 reconstructed time-of-flight series of spectral frames (colour scale as in **c**). In the X-ray-first region, note the strong suppression of the central peak due to N_2^{++} dications. The effect of impulsive alignment is evident in the infrared-first regime.

pulse. This experiment is typical of the class of stroboscopic pump-probe approaches that are widely used to investigate the temporal evolution of ultrafast processes¹⁷. Because such phenomena can be heavily obscured by timing uncertainty, the dataset used here represents a critical test of our data-analytic approach.

Below, we show that our analysis successfully extracts known ultrafast phenomena, including the molecular vibrations of the N_2 system with periods as short as 15 fs. We also pinpoint the start and end of the infrared pulse to about 1 fs, reveal enhanced molecular dissociation during the infrared pulse, and report anticipated, but previously unobserved, wave packets excited by the X-ray pulse. These results do not depend sensitively on the specific values used for the algorithmic parameters; see Supplementary Information section 11. The concatenation window c is the most important parameter; the quality of the results improves steadily with increasing c until the concatenation window is comparable with the FWHM of the timing uncertainty.

Turning to the analysis, each of the 10^5 spectral snapshots including regions around mass-to-charge ratios of 7 and 14 was represented as a vector, with components consisting of the signal recorded in the pixels of the time-of-flight spectrum. The vectors were ordered according to the jitter-corrupted experimental time stamps, and concatenated to form supervectors. Because the jitter substantially exceeded the 50-as average interval between successive snapshots, the sequence of vectors was strongly compromised, and the concatenation order purely statistical. The results reported below were obtained with a 5,800-fold concatenation window ($c = 5,800$) spanning 290 fs. The diffusion map algorithm⁵ was used to investigate the intrinsic structure of the concatenated data. The resulting manifold was five-dimensional (as determined by the procedure outlined in ref. 18) and nonlinear. This data structure (Supplementary Fig. 9) precludes analysis using standard linear-algebraic means such as SVD; see Supplementary Information section 14.

NLSA of the experimental data reveals the presence of up to six modes with singular values above the noise plateau (Supplementary Fig. 10). As in standard SVD, the first topogram constitutes the mean, with the subsequent modes representing the various deviations from it. Each chronogram shows the time evolution of its respective topogram in the X-ray-first, infrared-first and overlap regimes. Chronograms describing the time evolution of the signal and of the time-of-flight spectra for the first four modes are shown in Figs 1 and 2.

As in standard SVD, unless the physical processes at work are independent and non-degenerate, a single mode need not represent a complete physical process. Therefore, extraction of the individual physical processes requires additional information. However, the measured behaviour of the system is a linear combination of the modes obtained by data analysis. Thus, the features revealed by each mode constitute key elements of the processes at work and provide insight into the behaviour of the system.

With the above caveat in mind, we now discuss the key features of each of the modes obtained by our analysis (Figs 1 and 2). In the infrared-first regime, the modes reveal well-known features associated with the impulsive orientational alignment of N_2 molecules, with successive modes capturing the average and higher moments of the aligned distribution¹⁶.

In all modes, clear features mark the time span during which the infrared and X-ray pulses overlap. The sharp turning points flanking the overlap region in mode 3 (Fig. 3) are separated by 36 ± 2 fs. We associate this time span with the period during which the infrared pulse was sufficiently intense to affect the process reflected in this mode in the overlap regime.

We now turn to specific features in modes 3 and 4 (Fig. 3), which shed light on the behaviour of the system in the overlap regime, where the infrared pulse is active. Mode 3 concerns the detection of N_2^{++} , while mode 4 reveals the collection of two N^+ ions, one ejected towards

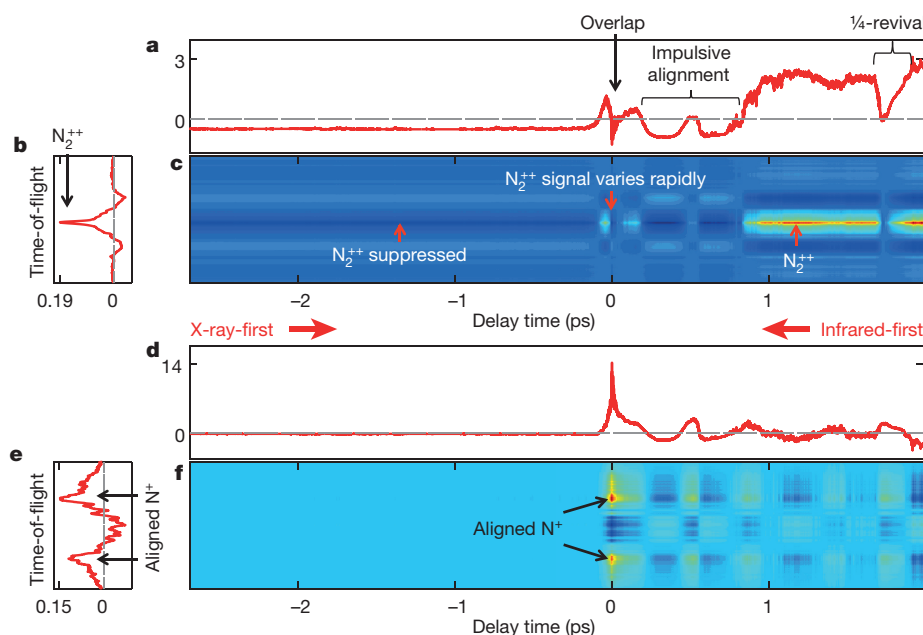


Figure 2 | Coulomb explosion of N_2 , singular modes 3 and 4. **a**, Mode-3 chronogram showing the time evolution of the mode. The vertical axis shows the number of molecular fragments reaching the detector in arbitrary units. Note the three different regimes: X-ray-first, overlap and infrared-first. A sharp feature marks the overlap region, where the X-ray and infrared pulses overlap. **b**, Mode-3 time-of-flight spectrum showing the region around a mass-to-charge ratio of 14 (the horizontal axis shows the signal strength in arbitrary units). The sharp central peak stems from N_2^{++} . **c**, Mode-3 reconstructed time-of-flight series of spectral frames (the colour scale indicates the signal strength: lowest signal, dark blue; highest

signal, red). The N_2^{++} signal is suppressed in the X-ray-first region, and varies rapidly in the overlap region. **d**, Mode-4 chronogram showing the time evolution of the mode; vertical axis as in **a**. A sharp feature marks the overlap region, in which the X-ray arrives while the infrared pulse is active. **e**, Mode-4 time-of-flight spectrum for mass-to-charge ratio of 14 (horizontal axis as in **b**). **f**, Mode-4 reconstructed time-of-flight series of spectral frames (colour scale as in **c**). A weak signal in the X-ray-first regime is invisible on this scale, but is clearly seen in Fig. 4. In the overlap region, the dissociation of molecules aligned with the polarization vector of the infrared pulse produces the two red peaks.

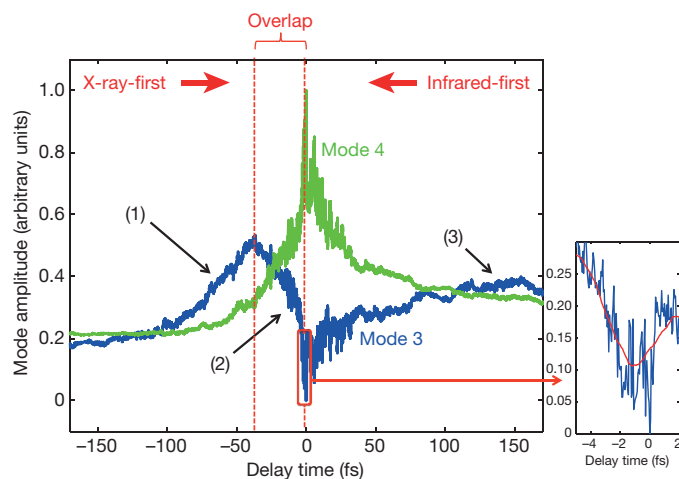


Figure 3 | Infrared/X-ray overlap region as revealed by singular modes 3 and 4. Sharp extrema in mode 3 (N_2^{++} dication signal) mark the start and end of the infrared pulse. The inset demonstrates that the extrema can be located to within ± 1 fs. The pulse width is thus 36 ± 2 fs at the intensity needed to strongly suppress the dication signal. The anticorrelation between modes 3 and 4 in the overlap region indicates that the corresponding N_2^{++} and N^+ channels are competing. The black arrows indicate: (1) dissociation of X-ray-generated N_2^{++} dications by the subsequent infrared pulse; (2) N_2^{++} suppression by the infrared pulse; and (3) build-up of infrared-generated molecular ions.

the detector and one ejected away from it. In the overlap region, modes 3 and 4 are strongly anticorrelated (with a correlation coefficient of -0.9917), indicating they are competing dissociation channels. It is tempting to associate the enhanced ejection of N^+ towards and away from the detector with enhanced dissociation of N_2^{++} along the polarization vector of the infrared pulse (Fig. 2f), as observed in strong-field experiments using optical pump and probe pulses under precisely controlled conditions^{19–22}. But the detector geometry used to obtain our experimental data strongly suppresses the detection of dissociation fragments ejected perpendicular to the detector axis. For this reason, we cannot reach a definitive conclusion on whether the dissociation is preferentially enhanced along the electric-field vector of the infrared pulse. Our results nonetheless highlight the type of detailed information yielded by our approach.

A key finding from our analysis concerns the observation of wave-packet dynamics in the X-ray-first regime (Fig. 4). In the majority of cases, the X-ray pulse creates core holes, which decay rapidly to form molecular dication states²³. In a minority of X-ray absorption events, a valence electron is ejected to produce N_2^+ ions. Abrupt events are expected to launch vibrational^{17,24,25} and/or charge²⁶ wave-packet dynamics. As shown in Fig. 4a, b, our analysis clearly reveals the presence of wave-packet oscillations and their revival in the X-ray-first regime. The periods of these oscillations (Fig. 4c, d) coincide closely with the known vibrations of the N_2^+ and N_2^{++} systems^{25,27}. However, only the 15-fs oscillation has been previously accessed in the time domain²⁵, with the other oscillation periods deduced from spectroscopic measurements²⁷. We also observe oscillations outside the well-studied 40–70-THz range (Supplementary Fig. 7 and Supplementary Information section 8). Initial results from a quantum-mechanical calculation corroborate the spectral features we extract data-analytically in the 10–70-THz frequency range (Supplementary Information section 9).

In the absence of a well-characterized physical process with a temporal knife-edge of sufficient abruptness, it is not straightforward to determine the exact time resolution achieved by our approach. Our results identify the start and end of the optical pulse, each to within about 1 fs (Fig. 3 inset). Given the width of the X-ray pulse (< 10 fs, possibly shorter than 6 fs; ref. 1), it seems surprising that such precise information can be obtained. Other things being equal, the time

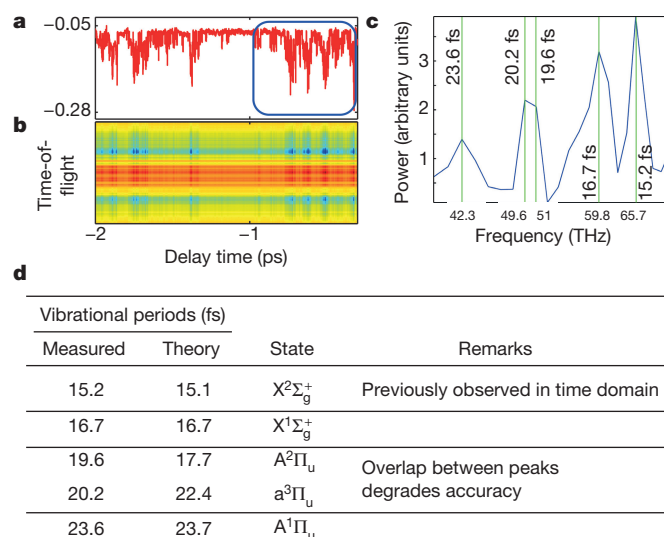


Figure 4 | Vibrational wave packets, singular mode 4, X-ray-first region. a, Chronogram in the X-ray-first region, showing the time evolution of a wave packet and its revival (the vertical axis shows the signal strength in the corresponding topogram in arbitrary units). b, Reconstructed time-of-flight series of spectral frames in the X-ray-first region (the colour scale indicates the signal strength: lowest signal, dark blue; highest signal, red). c, Fourier spectrum (blue) of the chronogram in the 690-fs-long boxed region in a, showing the frequency and period (green lines and labels) of each component. d, Comparison of observed periods with known vibrational modes of N_2 .

resolution is determined by the signal expected from an infinitely short pulse, convolved with the actual probe envelope. In agreement with previous work on partially coherent optical pulses²⁸, our results suggest that the X-ray pulse contains spikes that are sufficiently narrow to allow extraction of information beyond the nominal pulse envelope. The time resolution also depends on several other experimental parameters, including the timing uncertainty, the magnitude and uniformity of the time interval between snapshots²⁹, their signal-to-noise ratio, and the characteristics of the signal itself. Using modest computing resources (see Supplementary Information section 15), the present demonstration was achieved with 10^5 spectral snapshots with a signal-to-noise ratio of approximately 0.16, covering a time span approximately 17 times the FWHM of the timing uncertainty. Finally, the application of NLSA is essential even when the system under consideration is intrinsically linear. This stems from the large size of the matrix of concatenated data—containing about 10^{12} elements for the experimental data treated here—which greatly exceeds that amenable to standard SVD.

In summary, we have demonstrated a purely data-analytical approach that is capable of extracting the evolution and dynamics of complex systems from noisy snapshots on timescales much shorter than the uncertainty with which the data were recorded. We expect our approach to have a broad impact in many areas of science and technology; examples include geology and climate science, where timing of events can be uncertain, chemistry and biology, where reaction initiation can be non-uniform across a sample, and signal processing, where noise and timing jitter are prevalent.

Received 13 August 2015; accepted 18 February 2016.

- Glowina, J. M. *et al.* Time-resolved pump-probe experiments at the LCLS. *Opt. Express* **18**, 17620–17630 (2010).
- Gahl, C. *et al.* A femtosecond X-ray/optical cross-correlator. *Nature Photon.* **2**, 165–169 (2008).
- Löhl, F. *et al.* Electron bunch timing with femtosecond precision in a superconducting free-electron laser. *Phys. Rev. Lett.* **104**, 144801 (2010).
- Harmand, M. *et al.* Achieving few-femtosecond time-sorting at hard X-ray free-electron lasers. *Nature Photon.* **7**, 215–218 (2013).
- Coifman, R. R. *et al.* Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc. Natl Acad. Sci. USA* **102**, 7426–7431 (2005).

6. Giannakis, D. & Majda, A. J. Nonlinear Laplacian spectral analysis for time series with intermittency and low-frequency variability. *Proc. Natl Acad. Sci. USA* **109**, 2222–2227 (2012).
7. Berry, T., Cressman, R., Gregurić-Ferenček, Z. & Sauer, T. Time-scale separation from diffusion-mapped delay coordinates. *SIAM J. Appl. Dyn. Syst.* **12**, 618–649 (2013).
8. Aubry, N., Guyonnet, R. & Lima, R. Spatiotemporal analysis of complex signals: theory and applications. *J. Stat. Phys.* **64**, 683–739 (1991).
9. Packard, N., Crutchfield, J., Farmer, J. & Shaw, R. Geometry from a time series. *Phys. Rev. Lett.* **45**, 712–716 (1980).
10. Takens, F. in *Dynamical Systems and Turbulence, Warwick 1980* (eds Rand, D. A. & Young, L.-S.) Vol. 898 *Lecture Notes in Mathematics* (eds Dold, A. & Eckmann, B.) 366–381 (Springer, 1981).
11. Sauer, T., Yorke, J. A. & Casdagli, M. Embedology. *J. Stat. Phys.* **65**, 579–616 (1991).
12. Giannakis, D. & Majda, A. J. Comparing low-frequency and intermittent variability in comprehensive climate models through nonlinear Laplacian spectral analysis. *Geophys. Res. Lett.* **39**, L10710 (2012).
13. Dashti, A. *et al.* Trajectories of the ribosome as a Brownian nanomachine. *Proc. Natl Acad. Sci. USA* **111**, 17492–17497 (2014).
14. Stark, J., Broomhead, D. S., Davies, M. E. & Huke, J. Delay embeddings for forces systems. II. Stochastic forcing. *J. Nonlinear Sci.* **13**, 519–577 (2003).
15. Pellegrini, C. & Stohr, J. X-ray free-electron lasers—principles, properties and applications. *Nucl. Instrum. Methods Phys. Res. A* **500**, 33–40 (2003).
16. Stapelfeldt, H. & Seideman, T. *Colloquium: aligning molecules with strong laser pulses.* *Rev. Mod. Phys.* **75**, 543–557 (2003).
17. Rudenko, A. *et al.* Real-time observation of vibrational revival in the fastest molecular system. *Chem. Phys.* **329**, 193–202 (2006).
18. Coifman, R. R., Shkolnisky, Y., Sigworth, F. J. & Singer, A. Graph Laplacian tomography from unknown random projections. *IEEE Trans. Image Process.* **17**, 1891–1899 (2008).
19. Schmidt, M. *et al.* Fragment-emission patterns from the Coulomb explosion of diatomic molecules in intense laser fields. *Phys. Rev. A* **60**, 4706–4714 (1999).
20. Voss, S. *et al.* High resolution kinetic energy release spectra and angular distributions from double ionization of nitrogen and oxygen by short laser pulses. *J. Phys. B* **37**, 4239–4257 (2004).
21. Pavičić, D., Lee, K. F., Rayner, D. M., Corkum, P. B. & Villeneuve, D. M. Direct measurement of the angular dependence of ionization for N₂, O₂, and CO₂ in intense laser fields. *Phys. Rev. Lett.* **98**, 243001 (2007).
22. Guo, W., Zhu, J., Wang, B., Wang, Y. & Wang, L. Angular distributions of fragment ions of N₂ in a femtosecond laser field. *Phys. Rev. A* **77**, 033415 (2008).
23. Eberhardt, W., Stohr, J., Feldhaus, J., Plummer, E. W. & Sette, F. Correlation between electron emission and fragmentation into ions following soft-X-ray excitation of the nitrogen molecule. *Phys. Rev. Lett.* **51**, 2370–2373 (1983).
24. Bocharova, I. A. *et al.* Time-resolved Coulomb-explosion imaging of nuclear wave-packet dynamics induced in diatomic molecules by intense few-cycle laser pulses. *Phys. Rev. A* **83**, 013417 (2011).
25. De, S. *et al.* Following dynamic nuclear wave packets in N₂, O₂, and CO with few-cycle infrared pulses. *Phys. Rev. A* **84**, 043410 (2011).
26. Timmers, H. *et al.* Coherent electron hole dynamics near a conical intersection. *Phys. Rev. Lett.* **113**, 113003 (2014).
27. Dawber, G. *et al.* Threshold photoelectrons coincidence spectroscopy of doubly-charged ions of nitrogen, carbon monoxide, nitric oxide and oxygen. *J. Phys. B* **27**, 2191–2209 (1994).
28. Meyer, K. *et al.* Noisy optical pulses enhance the temporal resolution of pump-probe spectroscopy. *Phys. Rev. Lett.* **108**, 098302 (2012).
29. Jerri, A. J. The Shannon sampling theorem—its various extensions and applications: a tutorial review. *Proc. IEEE* **65**, 1565–1596 (1977).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank P. Bucksbaum, J. M. Glowacki and A. Natan for experimental data and comments on an earlier version of the manuscript, and acknowledge discussions with A. Dashti, D. Giannakis, A. Hosseini-Zadeh, A. Rudenko, M. Schmidt and P. Schwander. The research conducted by A.O. and R.F. was supported by the US Department of Energy, Office of Science, Basic Energy Sciences under award DE-SC0002164 (algorithm design and development, and data analysis), and by the US National Science Foundation under awards STC 1231306 (numerical trial models) and 1551489 (underlying analytical models). The research conducted by T.S. and S.R. was supported by the US Department of Energy, Office of Science, Basic Energy Sciences under award DE-FG02-04ER15612. T.S. thanks the Hamburg Centre for Ultrafast Imaging for a Mildred Dresselhaus Visiting Professorship.

Author Contributions A.O. proposed the approach. R.F. and A.O. developed the algorithm architecture. R.F. implemented the algorithm and, together with A.O., obtained results from experimental data. A.O. and R.S. interpreted the experimental results. S.R. and T.S. performed simulations of impulsive molecular alignment and provided expert advice. A.M.H., O.V. and R.S. performed quantum-mechanical calculations.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.O. (Ourmazd@uwm.edu).

Quantum phases from competing short- and long-range interactions in an optical lattice

Renate Landig¹, Lorenz Hruby¹, Nishant Dogra¹, Manuele Landini¹, Rafael Mottl¹, Tobias Donner¹ & Tilman Esslinger¹

Insights into complex phenomena in quantum matter can be gained from simulation experiments with ultracold atoms, especially in cases where theoretical characterization is challenging. However, these experiments are mostly limited to short-range collisional interactions; recently observed perturbative effects of long-range interactions were too weak to reach new quantum phases^{1,2}. Here we experimentally realize a bosonic lattice model with competing short- and long-range interactions, and observe the appearance of four distinct quantum phases—a superfluid, a supersolid, a Mott insulator and a charge density wave. Our system is based on an atomic quantum gas trapped in an optical lattice inside a high-finesse optical cavity. The strength of the short-range on-site interactions is controlled by means of the optical lattice depth. The long (infinite)-range interaction potential is mediated by a vacuum mode of the cavity^{3,4} and is independently controlled by tuning the cavity resonance. When probing the phase transition between the Mott insulator and the charge density wave in real time, we observed a behaviour characteristic of a first-order phase transition. Our measurements have accessed a regime for quantum simulation of many-body systems where the physics is determined by the intricate competition between two different types of interactions and the zero point motion of the particles.

Experiments with cold atoms have contributed in many ways to the elucidation of the fundamental behaviour of quantum matter⁵. An example is the realization of the Bose–Hubbard model, where the balance between the kinetic energy of particles moving in an optical lattice and the on-site collisional interactions drives a quantum phase transition from a superfluid to a Mott insulating phase^{6,7}. While collisions between atoms are naturally present in quantum gases and give rise to short-range interactions⁸, longer-range interactions are more elusive. To investigate the latter, ultracold gases of particles with large magnetic or electric dipole moments^{9,10}, atoms in Rydberg states¹¹, or cavity-mediated interactions³ have been studied. Indeed, Hubbard models with additional nearest-neighbour interactions are already predicted to show intriguing phases, such as charge and spin density waves, supersolids, topological phases or checkerboard and stripe phases^{12–18}.

In our experiment, we achieve independent control over three energy scales by combining an optical lattice with cavity-mediated interactions (Fig. 1). The underlying static lattices along all three directions are necessary to study the direct competition between short- and long-range interactions, different from the situation very recently investigated in ref. 19. Aspects of this scenario, in which on-site interactions compete with infinite-range interactions, have been theoretically studied in the context of self-consistent extended Hubbard models and various phases have been predicted^{20–22}. The starting point is a Bose–Einstein condensate (BEC) of $4.2(4) \times 10^4$ ⁸⁷Rb atoms which is prepared inside the ultrahigh-finesse optical cavity (here the number in parentheses gives the uncertainty on the final digit). The optical lattice is formed by three mutually orthogonal standing waves. The lattice along the *y* axis at wavelength $\lambda_y = 670$ nm splits the BEC into

a stack of about 60 weakly coupled two-dimensional (2D) layers. These 2D layers are then exposed to a square lattice in the *x*–*z* plane formed by one free space lattice and one intracavity optical standing wave, both at a wavelength of $\lambda = 785.3$ nm. They create periodic optical potentials of equal depths V_{2D} along both directions, which we will specify in units of the recoil energy $E_R = \hbar^2/2m\lambda^2$, where *m* denotes the mass of ⁸⁷Rb. In addition to the lattice potential, the atoms are exposed to an overall harmonic confinement, which results in a maximum density of 2.8 atoms per lattice site at the centre of the trap. The standing wave along the *z* axis fulfils a second role as it controls long-range interactions via off-resonant scattering into the optical resonator mode. The photons are scattered off the trapped atoms and are delocalized within the cavity mode, thereby mediating atom–atom interactions of infinite range (see Methods). These infinite-range interactions create λ -periodic atomic density–density correlations on the underlying $\lambda/2$ -periodic square lattice⁴. The correlations can lead to the breaking of a \mathbb{Z}_2 -symmetry between the two checkerboard sublattices²³, defined by either even or odd sites, resulting in the appearance of a self-consistent optical potential with alternating strength.

In a wide range of the parameter space, the system can be described by a lattice model with long-range interactions (see Methods and Extended Data Fig. 1), given by:

$$\hat{H} = -t \sum_{\langle e,o \rangle} (\hat{b}_e^\dagger \hat{b}_o + \text{h.c.}) + \frac{U_s}{2} \sum_{i \in e,o} \hat{n}_i (\hat{n}_i - 1) - \frac{U_1}{K} \left(\sum_e \hat{n}_e - \sum_o \hat{n}_o \right)^2 - \sum_{i \in e,o} \mu_i \hat{n}_i \quad (1)$$

Here *e* and *o* denote all even and odd lattice sites respectively, \hat{b}_i (\hat{b}_i^\dagger) are the bosonic annihilation (creation) operators at site *i*, \hat{n}_i counts the number of atoms on site *i* and *K* is the total number of sites. The first term represents the tunnelling between neighbouring sites at rate *t* and favours delocalization of the atoms within a 2D layer, supporting superfluidity. The second term describes the on-site interaction with strength U_s controlled via V_{2D} . Its energy is minimized if the atomic wavefunctions are localized on individual lattice sites, with balanced populations on even and odd sites and vanishing spatial coherence. The infinite-range interactions are captured by the third term and favour, for positive U_1 , a particle imbalance between even and odd sites. This global atom–atom interaction strength U_1 is proportional to V_{2D} and inversely proportional to the detuning $\Delta_c = \omega_z - \omega_c$, where ω_z is the frequency of the *z*-lattice beam and ω_c is the cavity resonance frequency (see Methods). The last term represents the effective chemical potential $\mu_i = \mu - \epsilon_i$, incorporating the chemical potential μ and the external trapping potential ϵ_i on lattice site *i*. In the absence of long-range interactions, equation (1) reduces to the Bose–Hubbard model.

To explore the phase diagram of \hat{H} , the lattices along the *x* and *z* direction are simultaneously ramped up to a certain value V_{2D} , keeping the total ramp time constant. This procedure is repeated for different

¹Institute for Quantum Electronics, ETH Zurich, 8093 Zurich, Switzerland.

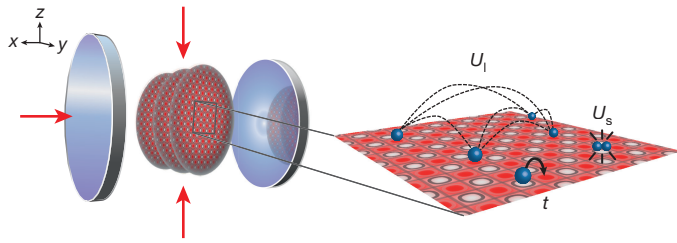


Figure 1 | Illustration of the experimental scheme that realizes a lattice model with on-site and infinite-range interactions. Left, a stack of 2D systems along the y axis is loaded into a 2D optical lattice (red arrows) between two mirrors (shown grey). The cavity induces atom-atom interactions of infinite range. Right, illustration of the competing energy scales: tunnelling t , on-site interactions U_s and long-range interactions U_l .

relative strengths of short- and long-range interactions (U_l/U_s), controlled via the detuning Δ_c .

To detect a superfluid–insulator phase transition, we probe the spatial coherence of the gas by turning off all confining potentials and taking absorption images of the atomic cloud after ballistic expansion. Figure 2a shows measured projected momentum distributions for four different V_{2D} , together with extracted vertical line sums. For small lattice depth V_{2D} , spatial coherence can be observed, characterized by a narrow momentum distribution of the cloud and a large BEC fraction f , extracted from a bimodal fit to the distribution. When increasing V_{2D} , the momentum distributions broaden, indicating a drop of coherence, and f reduces. We observe a kink in f as a function of the interaction strength U_s/t (for details, see Methods and Extended Data Fig. 4), which we associate with the formation of an insulating phase in

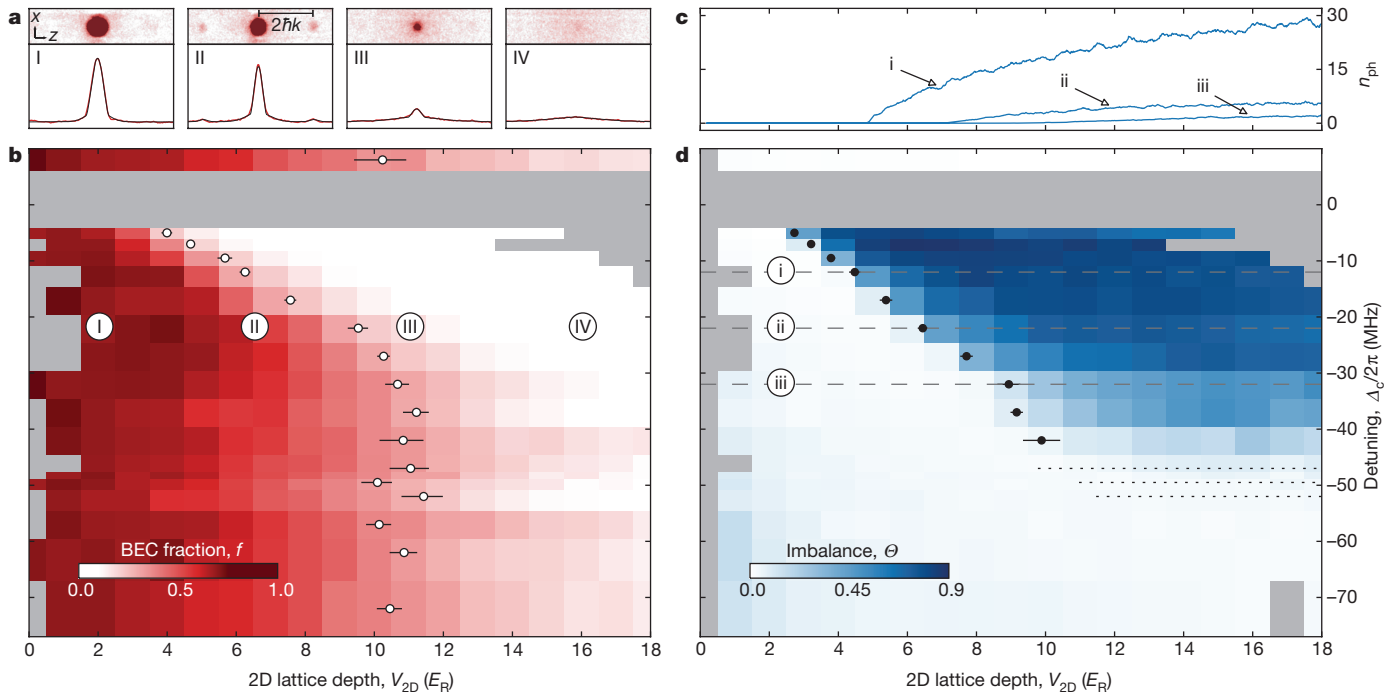


Figure 2 | Characterization of the phases. a–d, Characterization via spatial coherence (a, b) and via even-odd imbalance (c, d). a, Absorption images in the x - z plane (upper panels), and the same signal integrated along the cavity axis (lower panels, red), taken after a ballistic expansion for lattice depths V_{2D} of $2E_R$ (I), $6.5E_R$ (II), $11E_R$ (III) and $16E_R$ (IV) at $\Delta_c/2\pi = -22$ MHz. Black lines show fits with a bimodal distribution including higher momentum peaks. Owing to the cavity mirrors, the field of view along the x direction is restricted. b, Extracted BEC fraction f as a function of V_{2D} and Δ_c . White points mark the transition from a superfluid to an insulating phase and are obtained from a piecewise linear fit to the BEC fraction (see Extended Data Fig. 4). Error bars indicate fit uncertainties and contain contributions from the s.d. and from

the cloud and a loss of superfluidity²⁴. The extracted transition points are shown as white points in Fig. 2b. We confirmed that coherence between different lattice sites is restored when ramping down the 2D lattice potential again.

An even-odd imbalance causes a λ -periodic density modulation that acts as a Bragg grating, off which photons from the z -lattice beam are scattered into the cavity mode and vice versa. The amplitude of the scattered light field adiabatically follows the atomic density distribution⁴ and is continuously monitored using a heterodyne detection (see Methods). Figure 2c displays mean intracavity photon numbers n_{ph} measured as a function of V_{2D} . The onset of a cavity field is clearly visible and is taken as the transition point to a phase with even-odd imbalance Θ , marked with black points in Fig. 2d (for details, see Methods and Extended Data Fig. 4). The imbalance Θ can be quantified using equation (2) (see Methods):

$$\Theta = \left| \frac{\sum_e \langle \hat{n}_e \rangle - \sum_o \langle \hat{n}_o \rangle}{\sum_e \langle \hat{n}_e \rangle + \sum_o \langle \hat{n}_o \rangle} \right| \approx \frac{1}{N} \sqrt{n_{ph} \frac{\Delta_c^2}{\eta^2}} \quad (2)$$

Here, η is the two-photon Rabi frequency of the scattering process and N is the total atom number.

To establish a phase diagram, we combine all determined transition points in Fig. 3. We identify four phases that arise from the competition of the three energy scales: a superfluid (SF), a supersolid (SS), a Mott insulator (MI) and a charge density wave (CDW) phase. Far away from cavity resonance, that is, $\Delta_c/2\pi \lesssim -52$ MHz, U_l becomes small and the system undergoes, for large enough V_{2D} , a transition from an SF to an MI phase. The latter is characterized by a loss of coherence,

the stability of the fit (see Methods). c, Scattered photons n_{ph} of single repetitions as a function of V_{2D} for pump-cavity detunings $\Delta_c/2\pi$ of -12 MHz (i), -22 MHz (ii) and -32 MHz (iii). d, Imbalance Θ mapped as a function of Δ_c and V_{2D} . We assign the onset of a scattered cavity light field (black points) to the formation of a phase with even-odd imbalance. In the region indicated by the three dotted lines at values $\Delta_c/2\pi = \{-47, -49.5, -52\}$ MHz, the onset of the cavity light field showed a large variation. Error bars indicate the s.d. of the fit; an additional systematic error of $0.2E_R$ stems from the data analysis. The detection background is growing with decreasing V_{2D} and increasing detuning from cavity resonance (see Methods). Grey areas were not recorded.

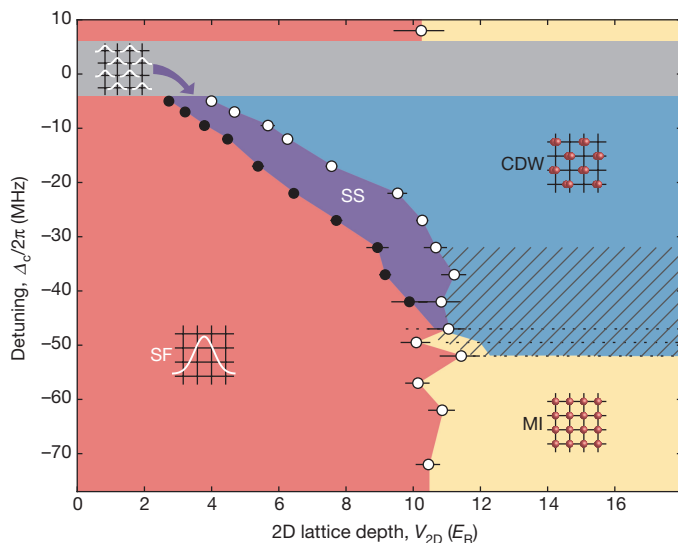


Figure 3 | Phase diagram. The four phases are indicated by different colours: SF (red), SS (violet), CDW (blue) and MI (yellow). Simplified density distributions are schematically illustrated for the homogeneous case with, on average, one atom per site. Data points (from Fig. 2b, d) show the experimentally obtained phase transition points recorded for increasing V_{2D} : Black data points indicate the onset of an even-odd imbalance, white data points depict where spatial coherence is lost. Increasing the 2D lattice depth V_{2D} simultaneously increases short- and long-range interactions. The detuning Δ_c changes only the strength of the long-range interactions. The slanted lines indicate the region where CDW and MI phases may coexist. At detuning $\Delta_c/2\pi = +8$ MHz, U_1 becomes negative and favours zero imbalance, thus only SF and MI phases appear. No data were taken at detunings indicated by the grey bar. A version of the phase diagram in Hamiltonian parameters is shown in Extended Data Fig. 2.

as well as the absence of an even-odd imbalance. The observed SF to MI transition line is shifted to larger values of V_{2D} than theoretically expected for a homogeneous system²⁵, which we attribute to the harmonic confinement of our 2D systems²⁶. Approaching cavity resonance increases U_1 . Above $\Delta_c/2\pi \approx -52$ MHz, this leads to the formation of a structured phase with even-odd imbalance, heralded by the onset of a light field scattered into the cavity. Depending on the relative strength of tunnelling and short-range interactions, the structured phase can either be an SS phase, where superfluidity is supported, or a CDW phase, where spatial coherence is lost. The identification of the SS phase is further supported by the observation of additional interference peaks corresponding to a λ -periodic density modulation (see Extended Data Fig. 5)¹⁵. The SF to SS phase boundary shifts to smaller V_{2D} when approaching the cavity resonance³. The transition line from an SS to a CDW follows the same trend. We attribute the loss of coherence in the CDW phase to a reduced nearest-neighbour tunnelling, which has its origin in an energy offset between even and odd lattice sites. This energy offset is a result of the optical potential created by the interference of the field scattered into the cavity with the field of the z lattice, and is shown in Extended Data Fig. 6. Our experimental resolution currently does not allow us to assess the precise topology of the multicritical region.

For long-range interactions dominating over tunnelling and short-range interactions, we observe a maximum even-odd imbalance Θ above 0.9, implying that mostly even or odd sites are occupied²⁷. This imbalance is significantly lower between -32 MHz $\lesssim \Delta_c/2\pi \lesssim -52$ MHz, see Fig. 2d. A possible explanation is the coexistence of CDW and MI phases²¹, which is supported by the external trapping potential, making it energetically costly for the system to place atoms away from the trap centre. Conversely, a homogeneous system with non-integer filling would turn into a structured phase with even-odd imbalance for any finite long-range interaction

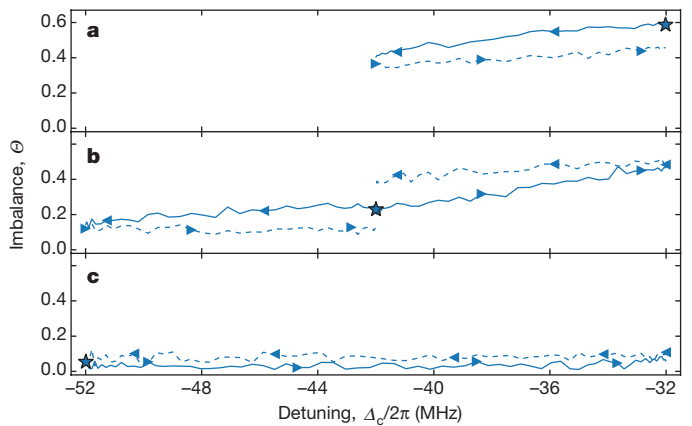


Figure 4 | Hysteretic behaviour of the CDW to MI transition. Shown is imbalance Θ recorded by varying $\Delta_c/2\pi$ at a rate of 0.67 MHz ms^{-1} , for fixed $V_{2D} = 14E_R$. The initial detunings $\Delta_c/2\pi$, indicated by stars, are -32 MHz (a), -42 MHz (b) and -52 MHz (c). Arrows signify the ramp directions; dashed lines show the return to the starting point. Curves are rescaled to take atom loss into account and contain three to nine averages, binned at $400 \mu\text{s}$ (see Methods).

(see Methods and Extended Data Fig. 3). Since the particles in the MI regions do not scatter into the cavity, the cavity field will rapidly vanish when the size of these regions increases. We conclude from the signal-to-noise ratio of our detection that below $\Delta_c/2\pi = -52$ MHz the technical noise does not allow us to detect an even-odd imbalance below 0.01.

We now study the evolution between the predominantly insulating CDW and MI phases. We initialize the system in the insulating region ($V_{2D} = 14E_R$) at a certain detuning Δ_c and then continuously vary U_1 by changing Δ_c , before returning to the initial value of Δ_c (see Methods). The cavity output field tracks the instantaneous even-odd imbalance Θ in real time. Figure 4a shows the evolution of the imbalance when decreasing Δ_c from an initial value in the CDW phase. The data show a hysteretic behaviour with a lower imbalance on return. The imbalance evolution for a starting value of Δ_c in the transition region between CDW and MI phases shows a similar behaviour when decreasing Δ_c and the opposite behaviour when increasing Δ_c . In Fig. 4c, where we started in the MI phase, the imbalance remains low throughout the measurement. We measured the hysteretic behaviour to be insensitive to the ramp speed (Extended Data Fig. 7).

This hysteretic behaviour of the system points towards a first-order phase transition between CDW and MI phases. When starting in an MI phase and increasing U_1/U_s beyond a certain point, the CDW phase will become energetically favourable, but cannot be reached because of an energy barrier between the two phases. Further increasing U_1/U_s lowers this energy barrier until the system is driven out of the metastable state. We suggest that this is activated in our inhomogeneous system by λ -periodic density-density correlations that are created in residual compressible regions, or superfluid shells, acting as impurities. In the opposite direction, moving from a CDW to an MI phase, the energy offset between even and odd lattice sites stabilizes the CDW phase beyond the point where the MI phase becomes energetically favourable, which results in the observed hysteretic behaviour.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 October 2015; accepted 5 February 2016.

Published online 11 April 2016.

1. Baier, S. et al. Extended Bose-Hubbard models with ultracold magnetic atoms. Preprint at <http://arXiv.org/abs/1507.03500> (2015).
2. Yan, B. et al. Observation of dipolar spin-exchange interactions with lattice-confined polar molecules. *Nature* **501**, 521–525 (2013).

3. Baumann, K., Guerlin, C., Brennecke, F. & Esslinger, T. Dicke quantum phase transition with a superfluid gas in an optical cavity. *Nature* **464**, 1301–1306 (2010).
4. Mottl, R. *et al.* Roton-type mode softening in a quantum gas with cavity-mediated long-range interactions. *Science* **336**, 1570–1573 (2012).
5. Bloch, I., Dalibard, J. & Nascimbène, S. Quantum simulations with ultracold quantum gases. *Nature Phys.* **8**, 267–276 (2012).
6. Greiner, M., Mandel, O., Esslinger, T., Hensch, T. W. & Bloch, I. Quantum phase transition from a superfluid to a Mott insulator in a gas of ultracold atoms. *Nature* **415**, 39–44 (2002).
7. Köhl, M., Moritz, H., Stöferle, T., Schori, C. & Esslinger, T. Superfluid to Mott insulator transition in one, two, and three dimensions. *J. Low Temp. Phys.* **138**, 635–644 (2005).
8. Weiner, J., Bagnato, V. S., Zilio, S. & Julienne, P. S. Experiments and theory in cold and ultracold collisions. *Rev. Mod. Phys.* **71**, 1–85 (1999).
9. Ni, K.-K. *et al.* A high phase-space-density gas of polar molecules. *Science* **322**, 231–235 (2008).
10. Stuhler, J. *et al.* Observation of dipole-dipole interaction in a degenerate quantum gas. *Phys. Rev. Lett.* **95**, 150406 (2005).
11. Heidemann, R. *et al.* Rydberg excitation of Bose-Einstein condensates. *Phys. Rev. Lett.* **100**, 033601 (2008).
12. Micnas, R., Ranninger, J. & Robaszkiewicz, S. Superconductivity in narrow-band systems with local nonretarded attractive interactions. *Rev. Mod. Phys.* **62**, 113–171 (1990).
13. Dutta, O. *et al.* Non-standard Hubbard models in optical lattices: a review. *Rep. Prog. Phys.* **78**, 066001 (2015).
14. Góral, K., Santos, L. & Lewenstein, M. Quantum phases of dipolar bosons in optical lattices. *Phys. Rev. Lett.* **88**, 170406 (2002).
15. Kovrizhin, D. L., Pai, G. V. & Sinha, S. Density wave and supersolid phases of correlated bosons in an optical lattice. *Europhys. Lett.* **72**, 162–168 (2005).
16. van Otterlo, A. *et al.* Quantum phase transitions of interacting bosons and the supersolid phase. *Phys. Rev. B* **52**, 16176–16186 (1995).
17. Scarola, V. W. & Sarma, S. D. Quantum phases of the extended Bose-Hubbard Hamiltonian: possibility of a supersolid state of cold atoms in optical lattices. *Phys. Rev. Lett.* **95**, 033003 (2005).
18. Dalla Torre, E. G., Berg, E. & Altman, E. Hidden order in 1D Bose insulators. *Phys. Rev. Lett.* **97**, 260401 (2006).
19. Klinder, J. *et al.* Observation of a superradiant Mott insulator in the Dicke-Hubbard model. *Phys. Rev. Lett.* **115**, 230403 (2015).
20. Ritsch, H., Domokos, P., Brennecke, F. & Esslinger, T. Cold atoms in cavity-generated dynamical optical potentials. *Rev. Mod. Phys.* **85**, 553–601 (2013).
21. Li, Y., He, L. & Hofstetter, W. Lattice-supersolid phase of strongly correlated bosons in an optical cavity. *Phys. Rev. A* **87**, 051604 (2013).
22. Habibian, H., Winter, A., Paganelli, S., Rieger, H. & Morigi, G. Bose-glass phases of ultracold atoms due to cavity backaction. *Phys. Rev. Lett.* **110**, 075304 (2013).
23. Baumann, K., Mottl, R., Brennecke, F. & Esslinger, T. Exploring symmetry breaking at the Dicke quantum phase transition. *Phys. Rev. Lett.* **107**, 140402 (2011).
24. Jiménez-García, K. *et al.* Phases of a two-dimensional Bose gas in an optical lattice. *Phys. Rev. Lett.* **105**, 110401 (2010).
25. Krauth, W. & Trivedi, N. Mott and superfluid transitions in a strongly interacting lattice boson system. *Europhys. Lett.* **14**, 627–632 (1991).
26. Rigol, M., Batrouni, G. G., Rousseau, V. G. & Scalettar, R. T. State diagrams for harmonically trapped bosons in optical lattices. *Phys. Rev. A* **79**, 053605 (2009).
27. Caballero-Benitez, S. F. & Mekhov, I. B. Quantum optical lattices for emergent many-body phases of ultracold atoms. *Phys. Rev. Lett.* **115**, 243604 (2015).

Acknowledgements We thank U. Bissbort, G. Graf, S. Huber, G. Morigi, L. Pollet and H. Ritsch for discussions and F. Brennecke for contributions in the early design phase of the experiment. We acknowledge funding from Synthetic Quantum Many-Body Systems (a European Research Council advanced grant) and the EU Collaborative Project TherMiQ (Grant Agreement 618074), and also SBF support for Horizon2020 project QUIC and SNF support for NCCR QSIT and DACH project ‘Quantum Crystals of Matter and Light’.

Author Contributions R.L., L.H., N.D. and M.L. took the data and analysed them together with T.D. Contributions to the design of the experiment were made by R.M. All work was supervised by T.E. All authors contributed to discussions and the preparation of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.D. (donner@phys.ethz.ch).

METHODS

Preparation of a BEC in 2D layers. We produce a Bose–Einstein condensate (BEC) of $4.2(4) \times 10^4$ ^{87}Rb atoms at a temperature of $42(2)$ nK in the $|F, m_F\rangle = |1, -1\rangle$ hyperfine state where F and m_F are respectively the total angular momentum and the corresponding magnetic quantum number. The quantization axis is defined by a magnetic field pointing along the z direction. The BEC is confined to the centre of a TEM₀₀ mode of the cavity by an optical dipole trap at a wavelength of 852 nm, with trap frequencies of $\omega_{x,y,z}/2\pi = [70.6(3), 31.4(5), 29.4(2)]$ Hz. Further details of the cavity set-up can be found in ref. 3.

The trapped BEC is loaded into a blue-detuned optical lattice of wavelength $\lambda_y = 670$ nm oriented along the y direction. This is done by implementing a smooth amplitude ramp (S-ramp) in time t which is of the form: $V(t) = V_0[3(t/t_0)^2 - 2(t/t_0)^3]$, where V_0 is the final lattice depth and t_0 is the total duration of the ramp. The lattice depth is increased to a final value of $24.9(1) E_R^{670}$ in 100 ms where $E_R^{670} = h^2/2m\lambda_y^2$ is the atomic recoil energy with m being the mass of a ^{87}Rb atom. The trap frequencies are kept constant during the loading by increasing the dipole trap depth simultaneously with the blue-detuned lattice. In this way, the whole BEC is cut into roughly 60 2D layers with about 1,300 atoms in the central layer.

Loading into the square lattice. After the preparation of 2D layers, the BEC is exposed to a 2D optical lattice in the x – z plane at a wavelength of $\lambda = 785.3$ nm. The lattice along the z direction is formed by a free space retro-reflected standing wave laser field which is linearly polarized along the y direction. The lattice along the x direction is created by pumping the TEM₀₀ mode of the cavity with linear polarization along the z direction. The effect of interference between the x and z lattices on atoms is minimized by introducing a frequency offset of at least 5 MHz between the two laser frequencies. Both lattices are ramped simultaneously within a fixed time of 50 ms to a variable lattice depth V_{2D} again using the S-ramp. The lattice potential seen by the atoms is of the form: $V(x, z) = V_{2D}[\cos^2(kx) + \cos^2(kz)]$ where $k = 2\pi/\lambda$ is the wave number and V_{2D} is the depth of the lattice in units of the corresponding recoil energy $E_R = h^2/2m\lambda^2$.

Characterization of the optical lattices. The lattice depths along the y and z directions are calibrated using Raman-Nath diffraction²⁸, whereas the lattice depth along the x direction is calibrated using amplitude modulation spectroscopy between the lowest and the first two excited Bloch bands²⁹. We estimate the calibration uncertainties on all lattice depths to be smaller than 4%. The uncertainty in the intracavity optical lattice depth is enlarged to about 10% by shifts of the cavity resonance frequency due to atomic redistribution during the V_{2D} ramp and residual drifts of the input coupling into the resonator.

The heating effect of the near-resonant x – z optical lattices on the BEC is characterized by ramping back down the lattices after reaching the insulating regime. We recover a BEC fraction larger than 0.45 and observe an atom loss of 5–10%. Loading of lattices in the x – z plane also increases the overall confinement. The trap frequencies are $\omega'_{x,z}/2\pi = [170, 165]$ Hz at a typical lattice depth of $10E_R$.

Lattice model with long-range interactions. The single-particle Hamiltonian \hat{H}_{sp} , describing the dynamics of an atom strongly coupled to a single cavity mode and moving in a 2D layer in the presence of static optical lattices, is given as^{3,20}:

$$\hat{H}_{sp} = \hat{H}_0 + V_{\text{trap}}(x, z) + \hbar\eta(\hat{a}^\dagger + \hat{a})\cos(kx)\cos(kz) - \hbar(\Delta_c - U_0\cos^2(kx))\hat{a}^\dagger\hat{a} \quad (3)$$

\hat{H}_0 consists of the kinetic energy of the particle and the potential seen due to the optical lattices in the x – z plane:

$$\hat{H}_0 = \frac{\hat{p}_x^2}{2m} + \frac{\hat{p}_z^2}{2m} + V_{2D}(\cos^2(kx) + \cos^2(kz)) \quad (4)$$

$V_{\text{trap}}(x, z)$ incorporates the inhomogeneous confining potential seen by the atoms. \hat{a} (\hat{a}^\dagger) annihilates (creates) a photon in the cavity mode. Scattering of the light field from the z lattice into the cavity mode at a two-photon Rabi frequency η creates a self-consistent checkerboard lattice for the atoms and is represented by the third term in \hat{H}_{sp} . This term describes how the atomic motion self-consistently determines the occupation of the cavity field mode inducing infinite-range interactions between the atoms. The last term in \hat{H}_{sp} represents the cavity field in the rotating frame of the z lattice with $\Delta_c = \omega_z - \omega_c$. The effect of the dispersive shift of the cavity resonance frequency is also included with U_0 being the maximum light shift per atom.

The many-body description of the system is obtained by introducing the bosonic field operator $\hat{\psi}(\mathbf{r})$ ($\hat{\psi}^\dagger(\mathbf{r})$) which annihilates (creates) a particle at position

$\mathbf{r} = (x, z)$ and satisfies bosonic commutation relations. In the framework of second quantization, the many-body Hamiltonian \hat{H}^{2nd} reads:

$$\hat{H}^{2nd} = \int d\mathbf{r} \hat{\psi}^\dagger(\mathbf{r})[\hat{H}_{sp} + g_{2D}\hat{\psi}^\dagger(\mathbf{r})\hat{\psi}(\mathbf{r}) - \mu]\hat{\psi}(\mathbf{r}) \quad (5)$$

where μ is the chemical potential and g_{2D} is the modified short-range interaction strength in a 2D layer³⁰. We expand $\hat{\psi}(x, z)$ in the basis of Wannier functions localized on different lattice sites which are obtained from the lowest Bloch band defined by \hat{H}_0 :

$$\hat{\psi} = \sum_{\mathbf{m}} W_{\mathbf{m}}(x, z)\hat{b}_{\mathbf{m}} \quad (6)$$

where $\hat{b}_{\mathbf{m}}$ ($\hat{b}_{\mathbf{m}}^\dagger$) represent the annihilation (creation) operators of a single particle at site $\mathbf{m} = (m_x, m_z)\lambda/2$ and $W_{\mathbf{m}}(x, z)$ is the Wannier function localized on site \mathbf{m} . A site is referred to as even (odd) if $m_x + m_z$ is even (odd). The Wannier functions localized on neighbouring lattice sites are related to each other by a translation of the lattice constant. Keeping interactions only up to the nearest neighbouring sites, we obtain the Bose–Hubbard model with additional terms³¹

$$\begin{aligned} \hat{H}_{\text{wan}}^{2nd} = & -t \sum_{(e,o)} (\hat{b}_e^\dagger \hat{b}_o + \text{h.c.}) + \frac{U_s}{2} \sum_{i \in e,o} \hat{n}_i (\hat{n}_i - 1) \\ & + \hbar\eta M_0 (\hat{a}^\dagger + \hat{a}) \left(\sum_e \hat{n}_e - \sum_o \hat{n}_o \right) \\ & - \hbar(\Delta_c - \delta) \hat{a}^\dagger \hat{a} - \sum_{i \in e,o} \mu_i \hat{n}_i \end{aligned} \quad (7)$$

where t and U_s represent tunnelling and contact interaction in the Bose–Hubbard model³² and are defined as:

$$t = \iint dx dz W_i^*(x, z) \hat{H}_0 W_i(x, z - \lambda/2) \quad (8)$$

$$U_s = g_{2D} \iint dx dz |W_i(x, z)|^4 \quad (9)$$

$\mu_i = \mu - \epsilon_i$ describes the local chemical potential at site i incorporating the effect of V_{trap} and $\hat{n}_i = \hat{b}_i^\dagger \hat{b}_i$ counts the number of particles on site i . Indices e and o refer to all even and odd lattice sites, respectively. $\delta = U_0 M_1 N$ is the dispersive shift of the cavity due to the BEC with N being the total number of atoms. The two overlap integrals M_0, M_1 are defined as:

$$\begin{aligned} M_0 &= \iint dx dz W_i^*(x, z) \cos(kx) \cos(kz) W_i(x, z) \\ M_1 &= \iint dx dz W_i^*(x, z) \cos^2(kx) W_i(x, z) \end{aligned}$$

A higher-order correction to the tunnelling along the x direction by the self-consistent cavity lattice is neglected. The cavity decay rate κ is large compared to the atomic recoil frequency which allows us to adiabatically eliminate the cavity field³. Its steady state value is given by:

$$\hat{a} = \frac{\eta M_0}{\Delta_c - \delta + i\kappa} \left(\sum_e \hat{n}_e - \sum_o \hat{n}_o \right) \quad (10)$$

Hence the light leaking out of the cavity will be proportional to the imbalance of the number of atoms on the two kind of sites and it can herald the presence of a phase with broken \mathbb{Z}_2 -symmetry of the underlying static lattice. Inserting equation (10) into equation (7), we recover equation (1) of the main text, with cavity-mediated long-range interaction strength U_1 given by:

$$\begin{aligned} U_1 &= -K\hbar|\eta M_0|^2 \frac{\Delta_c - \delta}{(\Delta_c - \delta)^2 + \kappa^2} \\ &\stackrel{|\Delta_c| \gg \kappa, |\delta|}{\approx} -K\hbar|M_0|^2 \frac{\eta^2}{\Delta_c} \propto \frac{V_{2D}}{\Delta_c} \end{aligned} \quad (11)$$

To describe our stack of 2D layers within this theoretical framework, we assume that a system of many 2D layers can be combined to form one 2D layer with an accordingly larger number of lattice sites, containing all atoms.

Validity of the theoretical model. In the derivation of equation (1), we assume the validity of the single-band approximation. To deduce the experimental parameter space where this assumption holds, we compare the strength of all Hamiltonian

parameters with the excitation energy to the next higher Bloch band, as shown in Extended Data Fig. 1.

Phase diagram in terms of Hamiltonian parameters. We convert the phase diagram from Fig. 3 into Hamiltonian parameters (see Extended Data Fig. 2) using equations (8), (9) and (11), taking into account the effect of two nearly-degenerate polarization modes of the cavity in the definition of U_1 . Starting in an SF phase and increasing U_1/t takes the system into an SS phase and eventually into a CDW phase. At the transition from the SF to the SS phase, the system needs to overcome additional short-range interaction energy. As a result, an increasingly larger critical long-range interaction strength is required to enter the SS phase for increasing U_0/t . A similar effect is seen for the transition from an SS to a CDW phase.

For negligible tunnelling, a direct transition from an MI to a CDW phase is found at a relative strength of $U_1/U_0 = 0.66(4)$. In the absence of tunnelling and trapping potential the Hamiltonian (equation (1)) supports a stable MI only for commensurate filling. In this case, the phase boundary between MI and CDW lies at a relative strength of $U_1/U_0 = 0.5$. Deviations from this value can be attributed to the presence of the trap, incommensurate filling and the non-local nature of the long-range interactions.

For negligible U_1 , the transition from SF to MI is observed at a relative strength of $U_0/t = 28(4)$. The value is larger than the theoretically predicted value of $U_0/t \approx 16$ for a homogeneous system with unity filling²⁵, as discussed in the main text.

Effect of the trapping potential. The harmonic trapping potential experienced by the atoms has a stabilizing effect on the MI phase in the presence of long-range interactions. Extended Data Fig. 3 illustrates, for fixed atom number and zero tunnelling, the effect of the trapping potential in the presence of the self-consistent lattice potential. For any non-zero U_1 and for non-integer filling, the homogeneous system will arrange itself in a structured phase with even-odd imbalance. However, the presence of a trapping potential can favour the coexistence of MI and CDW phases or of an MI phase alone, since the system has to pay additional energy for arranging atoms away from the trap centre. This energy cost has to be compared with the gain in energy due to the formation of a CDW phase. For larger fillings, we expect that the system will develop a ‘wedding cake’ like structure, similar to experimental realizations of MI phases. In our system, the plateaus can also host partially modulated and fully modulated CDW phases. The presence of any CDW in the system will be signalled by a finite light field scattered into the cavity. We do not expect a qualitative change of the phase diagram that depends on the steepness of the trap.

Extraction of the BEC fraction. We take an absorption image of the atomic distribution in the x - z plane after 15 ms of ballistic expansion. The obtained momentum distribution is integrated over the cavity direction. We perform a bimodal fit to the resulting distribution, in which we distinguish two contributions. The first component represents coherent atoms diffracted by the lattice potential, captured by a Thomas–Fermi profile plus two Gaussian interference peaks at $\pm 2\hbar k$. The second component is a broad Gaussian distribution resulting from the incoherent addition of atomic signals from the insulating part of the cloud. The BEC fraction f is finally extracted from the ratio $f = N_c/N$, where N_c is the integrated atom number in the coherent part and N is the total atom number. N is obtained from the mean total atom number of all experimental data at low lattice depths, $V_{2D} \leq 2E_R$. For deeper lattices, the growing spatial extent of the incoherent background is affected by inhomogeneities in the imaging and by the cropped field of view.

To constrain the number of free fit parameters, we fix the position of the interference peaks with respect to the central peak. Furthermore, their widths are linearly correlated to the width of the central peak³³, see Fig. 2a. We double count the interference peaks to correct for the non-visible peaks along the x direction, where the field of view is cropped by the cavity mirrors. The contribution from the $\pm 2\hbar k$ peaks to the total atom number is of the order of a few per cent at most. The chequerboard lattice in the supersolid phase leads to extra interference peaks, which lie outside the field of view (see Extended Data Fig. 5). Their contribution to the overall atom number is even lower than the one from the $\pm 2\hbar k$ peaks in most parts of the phase diagram and is therefore neglected.

Extraction of the even-odd imbalance. During each experimental sequence, the Bragg scattered light leaking out of the cavity is detected with a heterodyne set-up³⁴ having a sensitivity of $0.67(1) V^2$ per intracavity photon. The heterodyne detection is insensitive to the laser field creating the static lattice along the cavity axis by the choice of orthogonal polarizations and a minimum frequency difference of 5 MHz. Both phase and magnitude of the light field are recorded. To separate out the coherent part of the light field, we apply a low pass filter to the quadratures before taking the absolute square to obtain an intracavity photon number n_{ph} . It is mapped to an even-odd particle imbalance obtained from equation (10) with $n_{ph} = \langle \hat{a}^\dagger \hat{a} \rangle$. We define the effective even-odd imbalance

Θ under the assumption of completely localized atoms on either even or odd sites ($M_0 = 1$):

$$\Theta = \frac{\left| \sum_e \langle \hat{n}_e \rangle - \sum_o \langle \hat{n}_o \rangle \right|}{\left| \sum_e \langle \hat{n}_e \rangle + \sum_o \langle \hat{n}_o \rangle \right|} = \frac{1}{N} \sqrt{n_{ph} \frac{\Delta_c^2}{\eta^2} \frac{1}{F(\Delta_c)}} \quad (12)$$

with

$$F(\Delta_c) = \sqrt{\frac{\Delta_c^2 \cos^2(\alpha)}{(\Delta_c - \delta - \delta_B/2)^2 + \kappa^2} + \frac{\Delta_c^2 \sin^2(\alpha)}{(\Delta_c - \delta + \delta_B/2)^2 + \kappa^2}} \Big|_{|\Delta_c| \gg \kappa, |\delta|, \delta_B} \approx 1 \quad (13)$$

describing the scattering into two linearly polarized TEM₀₀ eigenmodes of the cavity, separated owing to birefringence by $\delta_B = 2\pi \times 2.2$ MHz. Their polarizations are orthogonal and rotated by an angle $\alpha = 22^\circ$ with respect to the y - z axes. The cavity decay rate κ is $2\pi \times 1.25$ MHz and the effective two-photon Rabi frequency η for scattering into the two cavity modes is given by $\eta = 2\pi \times 2.7 \sqrt{V_{2D}} / \hbar \sqrt{\text{Hz}}$. Close to cavity resonance, the polarization of the Bragg scattered cavity field rotates slightly due to the birefringence, which we include in the detection efficiency of the heterodyne detection. The maximum dispersive shift U_0 per atom of each of the two cavity modes is $-2\pi \times 45.9$ Hz. The intracavity photon number is determined with a systematic uncertainty of 8%, leading to a relative uncertainty in Θ of less than 6%. The technical background level of the photodetection is converted into an imbalance background, which depends on Δ_c and V_{2D} . This causes the signal in the lower left corner of Fig. 2d. However, in the MI phase, we estimate from the s.d. of this background a resolution for Θ which is better than 1%.

Sample size. No statistical methods were used to predetermine sample size.

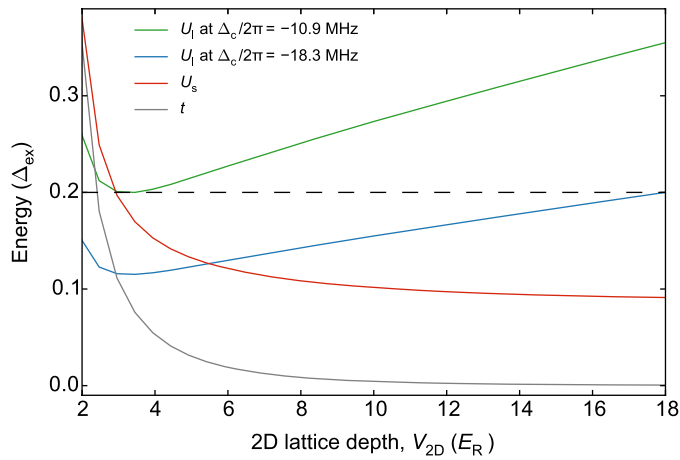
Phase boundaries. Coherence. We convert the 2D lattice depth V_{2D} to the corresponding ratio U_0/t of short-range interaction strength and nearest neighbour tunnelling by using the Wannier functions obtained from the lowest Bloch band of the applied static lattices. In this way, we obtain a BEC fraction f as a function of U_0/t (see Extended Data Fig. 4), which we fit with a piecewise linear function. The first kink in the fit is associated with the transition point to an insulating phase²⁴. By analysing the stability of the fit with respect to initial parameters, we deduce an additional uncertainty on top of the s.d. and include it in the error bar displayed for each transition point in Figs 2b and 3.

Even-odd imbalance. From each experimental repetition, we obtain a time trace of the light field scattered into the cavity. The maximum photon number $n_{ph,max}$ at the end of the trace and the corresponding lattice depth V_{2D} are averaged in a time window of 10 ms (spanning 7.8% of V_{2D}), resulting in one data point extracted per time trace. For each detuning, $n_{ph,max}(V_{2D})$ is fitted with a piecewise linear and power law function (see Extended Data Fig. 4) to determine the point where the intracavity light field starts building up. This method largely increases the signal to noise ratio while keeping systematic shifts of the onset point to below $0.2E_R$. In the region of $-52 \text{ MHz} \leq \Delta_c/2\pi \leq -47 \text{ MHz}$, the intracavity field becomes very small and fluctuates strongly from shot to shot (see Extended Data Fig. 4c). We therefore indicate a region for the transition to a phase with λ -periodic density modulation by dashed lines in Figs 2d and 3. The starting point of these dashed lines indicates the earliest onset of an even-odd imbalance including the s.d. of the fit. Owing to the fixed time of the lattice ramp, we cross the transition to an even-odd imbalanced phase non-adiabatically when ramping into deep lattices. The non-adiabaticity leads to a small shift of the onset point towards higher lattice depths, which can be seen in Fig. 2c. This behaviour was studied previously and explained with Kibble–Zurek theory²³. The described method of discretizing the data is intrinsically less sensitive to this type of shift compared to fitting a single time trace to extract the transition point.

Self-consistent chequerboard lattice. The SS and CDW phases give rise to a light field inside the cavity due to the Bragg scattering of z -lattice photons. This cavity field is self-consistent as it depends on the strength of the λ -periodic density modulation in the atomic cloud and has a depth $V_c = n_{ph} \times 12.3 \times 10^{-3} E_R$. Interference of this self-consistent x lattice with the field of the z lattice produces a chequerboard lattice potential of depth $V_{CB} = 2\sqrt{V_{2D}V_c}$, displayed in Extended Data Fig. 6. The line of constant V_{CB} bends towards smaller values of V_{2D} when approaching cavity resonance, substantiating the assumption that this energy offset causes the observed behaviour of the SS to CDW boundary line. When the energy offset between even and odd sites due to V_{CB} becomes comparable to the tunnelling energy, the effective tunnelling strength between nearest-neighbours reduces and higher-order tunnelling processes begin to play a significant role (U. Bissbort, personal communication).

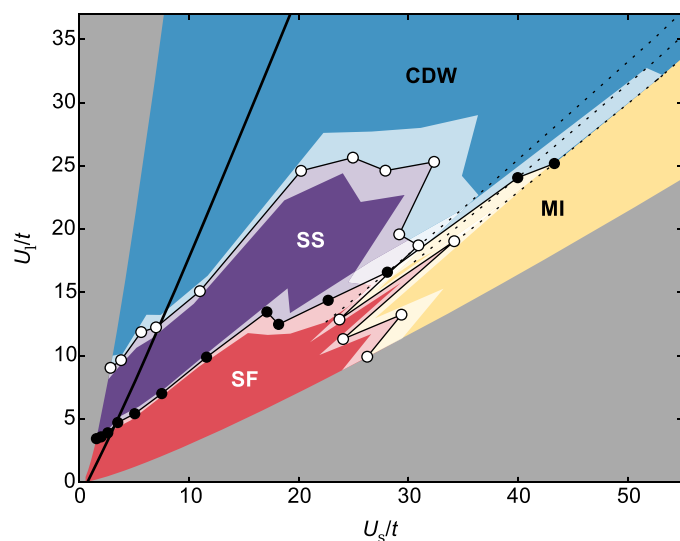
Hysteresis measurements. We initialize the system in the insulating region at either $V_{2D} = 14E_R$ or $18E_R$ using a 50-ms-long S-shaped amplitude ramp. The detuning $\Delta_c/2\pi$ is then changed with an S-shaped frequency ramp at an average speed of 0.67 MHz ms^{-1} reaching a different detuning value. After holding for 10 ms, we scan back to the initial detuning. Residual atom loss continuously reduces the measured mean intracavity photon number n_{ph} , which we take into account by rescaling the data before converting it into an imbalance Θ . The scaling factor is extracted from reference measurements, where we hold at different Δ_c for 50 ms. We deduce a linear decrease in n_{ph} by 48(4)% (41(4)%) for lattice depths of $V_{2D} = 14E_R$ ($18E_R$). After rescaling the data, we observe a remaining relative drift of the imbalance level of 8(4)% during the hold time. Extended Data Fig. 7 shows detuning scans performed at $V_{2D} = 18E_R$, where a similar hysteretic behaviour is observed as in Fig. 4 of the main text. To test the sensitivity of the hysteretic behaviour on the ramp speed, we slow down the frequency ramp by a factor of two and observe a comparable evolution of the even–odd imbalance; see Extended Data Fig. 7.

28. Morsch, O. & Oberthaler, M. Dynamics of Bose-Einstein condensates in optical lattices. *Rev. Mod. Phys.* **78**, 179–215 (2006).
29. Stöferle, T., Moritz, H., Schori, C., Köhl, M. & Esslinger, T. Transition from a strongly interacting 1D superfluid to a Mott insulator. *Phys. Rev. Lett.* **92**, 130403 (2004).
30. Petrov, D. S., Holzmann, M. & Shlyapnikov, G. V. Bose-Einstein condensation in quasi-2D trapped gases. *Phys. Rev. Lett.* **84**, 2551–2555 (2000).
31. Maschler, C., Mekhov, I. B. & Ritsch, H. Ultracold atoms in optical lattices generated by quantized light fields. *Euro. Phys. J. D* **46**, 545–560 (2008).
32. Jaksch, D., Bruder, C., Cirac, J. I., Gardiner, C. W. & Zoller, P. Cold bosonic atoms in optical lattices. *Phys. Rev. A* **81**, 3108–3111 (1998).
33. Spielman, I., Phillips, W. & Porto, J. Condensate fraction in a 2D Bose gas measured across the Mott-insulator transition. *Phys. Rev. Lett.* **100**, 120402 (2008).
34. Landig, R., Brennecke, F., Mottl, R., Donner, T. & Esslinger, T. Measuring the dynamic structure factor of a quantum gas undergoing a structural phase transition. *Nature Commun.* **6**, 7046 (2015).

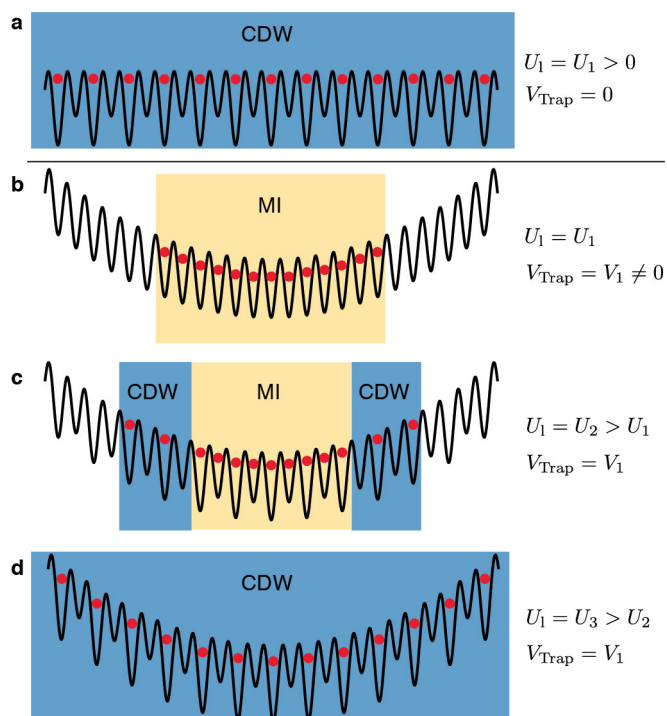


Extended Data Figure 1 | Validity of the single-band approximation.

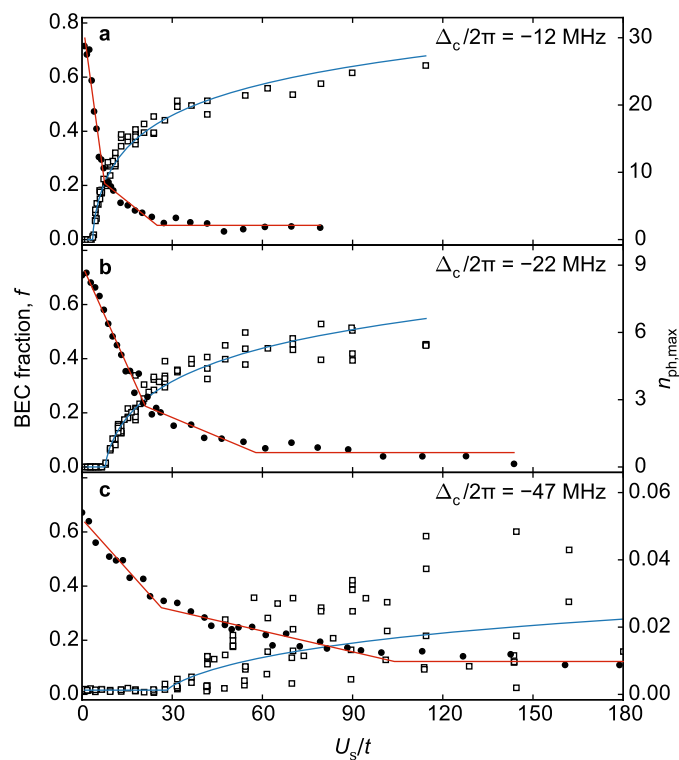
The energy scales of the Hamiltonian are plotted in units of the minimum gap Δ_{ex} between the lowest and the first excited Bloch band. The single-band approximation is assumed to be valid if all the energy scales (U_s , U_l and t) are at least 5 times smaller than Δ_{ex} , that is, if they lie below the black dashed line. This criterion is fulfilled for $\Delta_c/2\pi < -18.3$ MHz and $18E_R > V_{2D} > 3E_R$. For detunings in the interval $-18.3 \text{ MHz} < \Delta_c/2\pi < -10.9$ MHz, the approximation is only partially valid, depending on V_{2D} . We use this information to illustrate the region of validity in Extended Data Fig. 2.



Extended Data Figure 2 | Phase diagram plotted as a function of Hamiltonian parameters U_s/t and U_l/t . The experimental parameters in Fig. 3 have been converted to Hamiltonian parameters: the region of validity for this conversion lies to the right of the solid black line, grey areas were not recorded. The white data points indicate where spatial coherence is lost, and the black data points depict the onset of an even–odd imbalance. The white shaded regions around the data points represent the respective converted error bars. The dotted black lines show, as in Fig. 3, the region where the onset of the cavity light field showed a large variation.

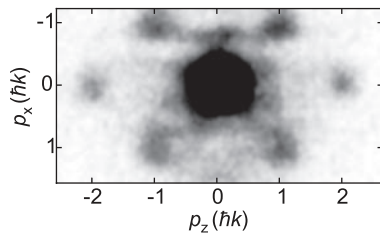


Extended Data Figure 3 | Influence of the trapping potential on the long-range interacting system. Shown are sketches of a 1D slice through a 2D layer, displaying the ground state configurations of 13 particles that depend on the relative influence of trapping potential and long-range interaction. **a**, Results for a homogeneous system with finite U_1 . **b–d**, The state of the system for increasing but finite U_1 , starting with small but finite U_1 (see legend at right).



Extended Data Figure 4 | Determination of the phase boundaries.

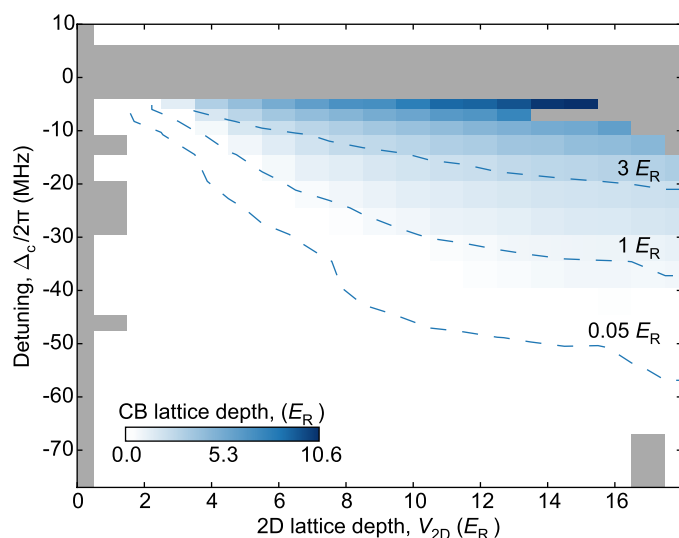
Shown are the BEC fraction f (averaged into 100 equally spaced bins) and maximum photon number $n_{\text{ph,max}}$ (filled and open symbols, respectively) as a function of U_s/t for detunings $\Delta_c/2\pi$ of -12 MHz (a), -22 MHz (b) and -47 MHz (c). The red curve in each panel shows the result of a piecewise linear fit to f . We confirmed that the initial BEC fraction has no systematic dependence on Δ_c . The blue curve displays a power law fit to $n_{\text{ph,max}}$.



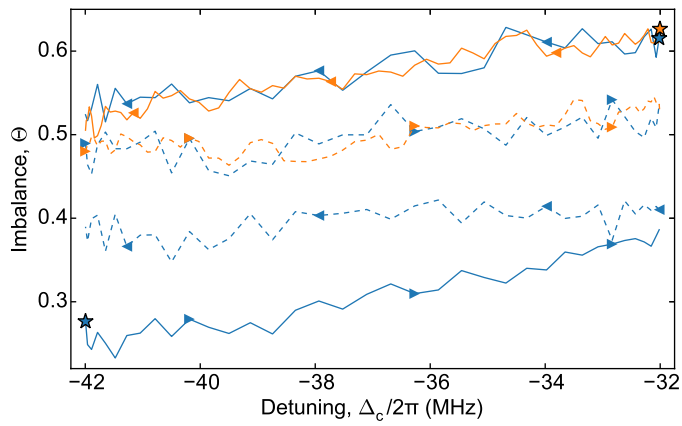
Extended Data Figure 5 | Momentum distribution in the SS phase.

Absorption image from a calibration measurement taken after a short ballistic expansion of 7 ms at a detuning of $\Delta_c/2\pi = -23$ MHz and a lattice depth V_{2D} 39% above the onset of an even–odd imbalance in the SS phase.

We observe interference peaks at $p_z = \pm 2\hbar k$. Additional interference peaks resulting from the emerging chequerboard lattice appear at $(p_x, p_z) = (\pm\hbar k, \pm\hbar k)$. This observation indicates an SS phase. These additional momentum peaks lie outside the field of view for the longer ballistic expansion time of 15 ms.



Extended Data Figure 6 | Strength of the self-consistent chequerboard lattice. The chequerboard (CB) lattice depth extracted from the measured mean intracavity photon number n_{ph} is shown as a function of the applied lattice depth V_{2D} and detuning Δ_c . The CB lattice depth becomes comparable to the depth of the static lattices close to cavity resonance, but drops rapidly when moving away due to its detuning dependence. Exemplary equipotential lines at $0.05 E_R$, $1 E_R$ and $3 E_R$ are shown.



Extended Data Figure 7 | Sensitivity to the ramp speed. The hysteretic behaviour in the insulating regime at $V_{2D} = 18E_R$ is shown. The detuning $\Delta_c/2\pi$ is ramped at two speeds, 0.67 MHz ms^{-1} (blue) and 0.33 MHz ms^{-1} (orange). Lines result from an average of two to five measurements, using $400 \mu\text{s}$ time bins. Stars signify starting points, arrows show the scan direction and dashed lines indicate the return to the starting point.

Nanocrack-regulated self-humidifying membranes

Chi Hoon Park^{1*†}, So Young Lee^{1*†}, Doo Sung Hwang^{1*†}, Dong Won Shin^{1†}, Doo Hee Cho¹, Kang Hyuck Lee¹, Tae-Woo Kim², Tae-Wuk Kim², Mokwon Lee³, Deok-Soo Kim³, Cara M. Doherty⁴, Aaron W. Thornton⁴, Anita J. Hill⁴, Michael D. Guiver^{5,6} & Young Moo Lee¹

The regulation of water content in polymeric membranes is important in a number of applications, such as reverse electrodialysis and proton-exchange fuel-cell membranes. External thermal and water management systems add both mass and size to systems, and so intrinsic mechanisms of retaining water and maintaining ionic transport^{1–3} in such membranes are particularly important for applications where small system size is important. For example, in proton-exchange membrane fuel cells, where water retention in the membrane is crucial for efficient transport of hydrated ions^{1,4–7}, by operating the cells at higher temperatures without external humidification, the membrane is self-humidified with water generated by electrochemical reactions^{5,8}. Here we report an alternative solution that does not rely on external regulation of water supply or high temperatures. Water content in hydrocarbon polymer membranes is regulated through nanometre-scale cracks ('nanocracks') in a hydrophobic surface coating. These cracks work as nanoscale valves to retard water desorption and to maintain ion conductivity in the membrane on dehumidification. Hydrocarbon fuel-cell membranes with surface nanocrack coatings operated at intermediate temperatures show improved electrochemical performance, and coated reverse-electrodialysis membranes show enhanced ionic selectivity with low bulk resistance.

Ion-exchange membranes are used in a wide range of applications for separations, energy conversion and energy storage systems, where selective barrier properties are essential for high performance in membrane-integrated systems. Membranes with selective transport surfaces, which have properties unlike those of the bulk material, have the potential to overcome the permeability and selectivity trade-off behaviour that is observed in many applications. For example, proton-exchange membrane (that is, cation-exchange membrane) research has focused on controlling membrane hydration for proton conduction under low-humidity or non-humidified conditions^{6,9}. In hydrocarbon proton-exchange membranes, water regulation has been achieved primarily through polymer architecture that induces phase-separated morphology between hydrophilic ion-conducting channels and the hydrophobic matrix, similar to state-of-the-art perfluorosulfonic acid proton-exchange membranes (such as Nafion), allowing high ion conduction with reduced overall membrane hydration (see also Supplementary Discussion 2.1)^{10–13}. Despite improvements in water retention using this approach, additional challenges have led to the performance being below expectations^{14,15}.

Here, we propose a new concept for regulating membrane hydration in low-humidity or non-humidified environments without modification of the morphology of an ion-exchange membrane, analogous to the water retention mechanisms of the cactus plant (such as *Ferocactus schwarzii*) (see Extended Data Fig. 1a, b). The cactus retains water by

opening and closing an array of stomatal openings, which respond to environmental conditions (see also Supplementary Discussion 2.2). To decrease water loss, stomata are open at night, during conditions of lower temperature and higher humidity. During the daytime, when hot and arid conditions prevail, the stomata are closed.

Figure 1a shows a conceptual diagram of a thin water-impermeable hydrophobic layer deposited on the membrane surface to regulate water exchange at the membrane surface (see also Supplementary Discussion 2.3). In this concept, the hydrophobic layer must resolve the paradox of conserving water within the bulk membrane while simultaneously not hindering ions that co-transport with water molecules through the surface of the membrane. Note that for an ion-conduction mechanism in which water is the transport medium, even very thin hydrophobic barriers can drastically reduce ion conductivity. Accordingly, to overcome this paradox, we deposited thin hydrophobic layers having very narrow water channels (nanocracks), which open under humidifying conditions. As shown in Fig. 1b and c, thin hydrophobic surface-coating layers are deposited by atmospheric plasma treatment^{16–18} (see Extended Data Figs 2a and b, 3b and c, 4a–e, and Supplementary Discussion 2.4), controlling the thickness and morphology pattern of the layers to allow water and ion molecules to pass through the coating. Dimensional swelling of the hydrated membranes induces reproducible and controllable nanocracked morphology patterns in the deposited hydrophobic surface layer, which then function as water channels (Fig. 1b and c, Extended Data Fig. 5a–f and Supplementary Discussion 2.5). Most importantly, we anticipated that the nanocracks would be partially closed under dry (that is, non-humidifying) conditions, thus acting as nanovalves for water and ions, allowing water conservation within the bulk membrane.

Using fuel-cell membranes to illustrate this concept, the electrochemical reaction generates water at the cathode, which can be absorbed through the water channel of the coated layer and humidifies the bulk membrane, thereby increasing the efficiency of the self-humidifying system. We used sulfonated poly(arylene ether sulfone) (BPSH) random copolymers as representative hydrocarbon proton-exchange membrane matrices for the hydrophobic atmospheric plasma treatment (Extended Data Fig. 2c). We chose BPSH because it has excellent chemical, thermal and mechanical stability but has less-developed phase-separated morphology¹⁹. BPSH40 and BPSH60, with molar ratios of disulfonated repeat units of 0.4 and 0.6, respectively, showed increased contact angles after hydrophobic coating. Atomic force microscopy (AFM) topology images of surface-coated BPSH membranes under low humidity show partially closed hydrophilic domains, whereas no morphological change appears in uncoated BPSH (Fig. 1b and Extended Data Fig. 3a). Also, we confirmed that there was a negligible difference in water channel

¹Department of Energy Engineering, College of Engineering, Hanyang University, Seoul 133-791, South Korea. ²Department of Life Science, College of Natural Science, Hanyang University, Seoul 133-791, South Korea. ³School of Mechanical Engineering, College of Engineering, Hanyang University, Seoul 133-791, South Korea. ⁴Manufacturing Flagship, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Clayton, Victoria 3168, Australia. ⁵State Key Laboratory of Engines, Tianjin University, Tianjin 300072, China. ⁶Collaborative Innovation Center of Chemical Science and Engineering (Tianjin), Tianjin 300072, China. [†]Present addresses: Department of Energy Engineering, Gyeongnam National University of Science and Technology, Jinju 660-758, South Korea (C.H.P.); Fuel Cell Research Center, Korea Institute of Science and Technology (KIST), 39-1 Hawolgok-dong, Seongbuk-gu, Seoul 136-791, South Korea (S.Y.L.); Department of Chemical Engineering, University of Illinois Chicago, Chicago, Illinois 60607, USA (D.S.H.); Fuel Cell Laboratory, Korea Institute of Energy Research, Daejeon 34129, South Korea (D.W.S.).

*These authors contributed equally to this work.

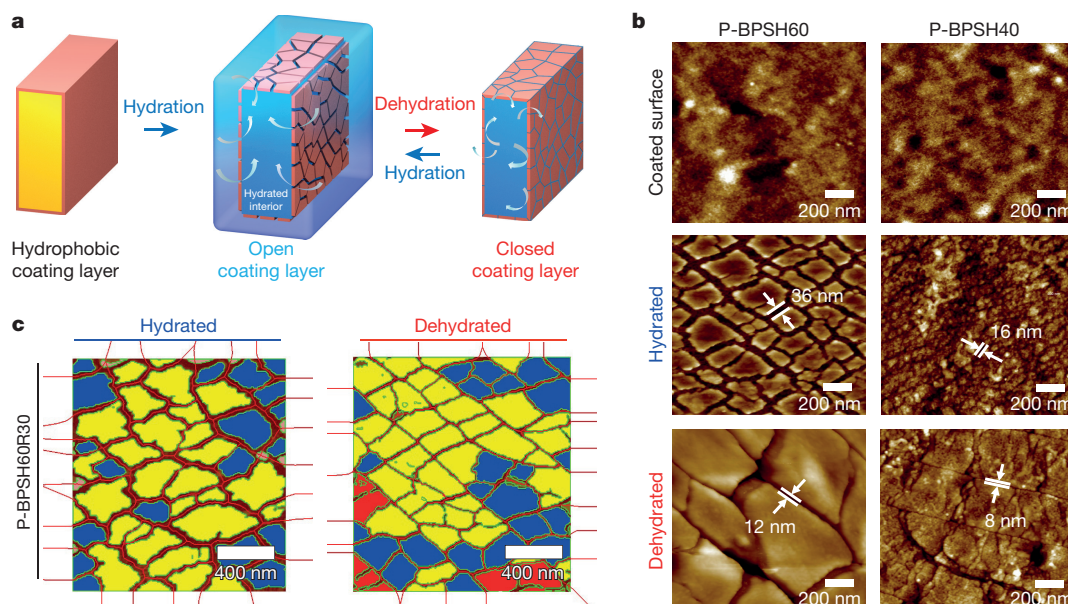


Figure 1 | Basic concept of self-humidifying nanovalved membrane.

a, A hydrophobic coating layer provides a self-controlled mechanism for water conservation using nanometre-sized cracks (nanocracks) tuned by membrane swelling behaviour in response to external humidity conditions, which act as nanovalves. **b**, AFM images reveal the self-controlled mechanism of plasma-coated membranes. Plasma-coated BPSH membranes have no visible nanocracks immediately after treatment (top panels). However, upon hydration in distilled water, membrane

swelling triggers the opening of the nanocracks, enabling water absorption (middle panels). During dehydration of plasma-coated membranes at 30% to 45% relative humidity (RH), the nanocracks become narrower, thus reducing water loss (bottom panels). **c**, Voronoi diagram analysis and tessellation entropy verified controllable nanocrack surface pattern images of plasma-coated membrane (P-BPSH60R30, where R30 indicates that the plasma treatment was repeated 30 times) in hydration (100% RH) and dehydration (30% to 45% RH).

size distribution inside the BPSH membranes before and after plasma treatment, as revealed by positron annihilation lifetime spectroscopy (PALS)²⁰ (see Fig. 2a). Their behaviour is consistent over the whole range of relative humidity (RH), demonstrating that water hydration properties and water channel morphology in the proton-exchange membrane matrix are not affected by the hydrophobic plasma surface coating.

To demonstrate the water-conserving effect, Fig. 2b and c shows that the water desorption rate for plasma surface-coated membranes (P-BPSH) is noticeably delayed compared with the uncoated membranes, and the water sorption rate is also delayed but to a lesser extent. This becomes more conspicuous after repeated sorption–desorption cycles, as shown by pulsatile dynamic vapour sorption (DVS). Despite the hydrophobic surface coating, the P-BPSH

membranes show amounts of bulk water within the membranes similar to the amounts within uncoated BPSH at each RH (see also Supplementary Table 1). From the DVS and PALS data, we establish that the bulk water channel morphology and the water diffusion coefficients within BPSH and P-BPSH are similar. However, a remarkable decrease in the initial water diffusion rates calculated from the initial slopes of each step in the DVS graphs occurs after hydrophobic surface coating, which retards the water sorption/desorption process. This behaviour is more substantial for desorption than for sorption, allowing the P-BPSH membranes to reduce water loss. We confirmed this experimental result theoretically by a mathematical water sorption–desorption model in Extended Data Fig. 6a–d (see also Supplementary Discussion 2.6).

To demonstrate this advantageous water-conserving effect using the hydrophobic membrane surface coating for hydrocarbon

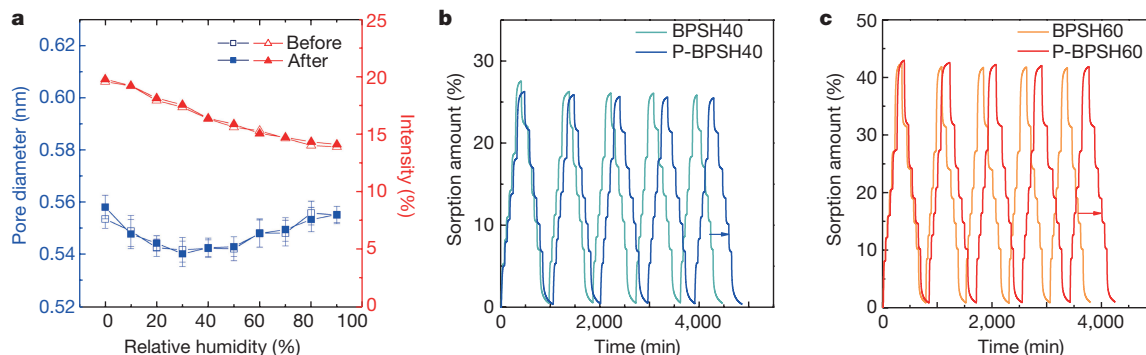


Figure 2 | Water-conserving effect of hydrophobic surface coating from retarded water desorption. **a**, PALS measures the membrane porosity (size and relative number or intensity of pores) and shows no difference in the membrane water transport pathways, as functions of RH, before and after surface coating of BPSH40 to produce P-BPSH40. Values are averages of at least five replicates; error bars represent 1 s.d. **b**, **c**, Pulsatile DVS measurement of BPSH40 (**b**) and BPSH60 (**c**) before surface coating (light blue and orange lines) and after surface coating (dark blue and red lines).

Here, dynamic hydration rates represented as sorption and desorption as well as water sorption amount (or water uptake) were measured at RH in stepwise increase from 0% to 90% and then decrease to 15% RH at 25 °C. Retardation of desorption was cumulative, although both membranes exhibited the same water uptake with similar hydration rate, as DVS cycles were repeated. The blue (**b**) and red (**c**) arrows indicate the delayed time of sorption or desorption for P-BPSH40 and P-BPSH60, respectively, during five DVS cycles.

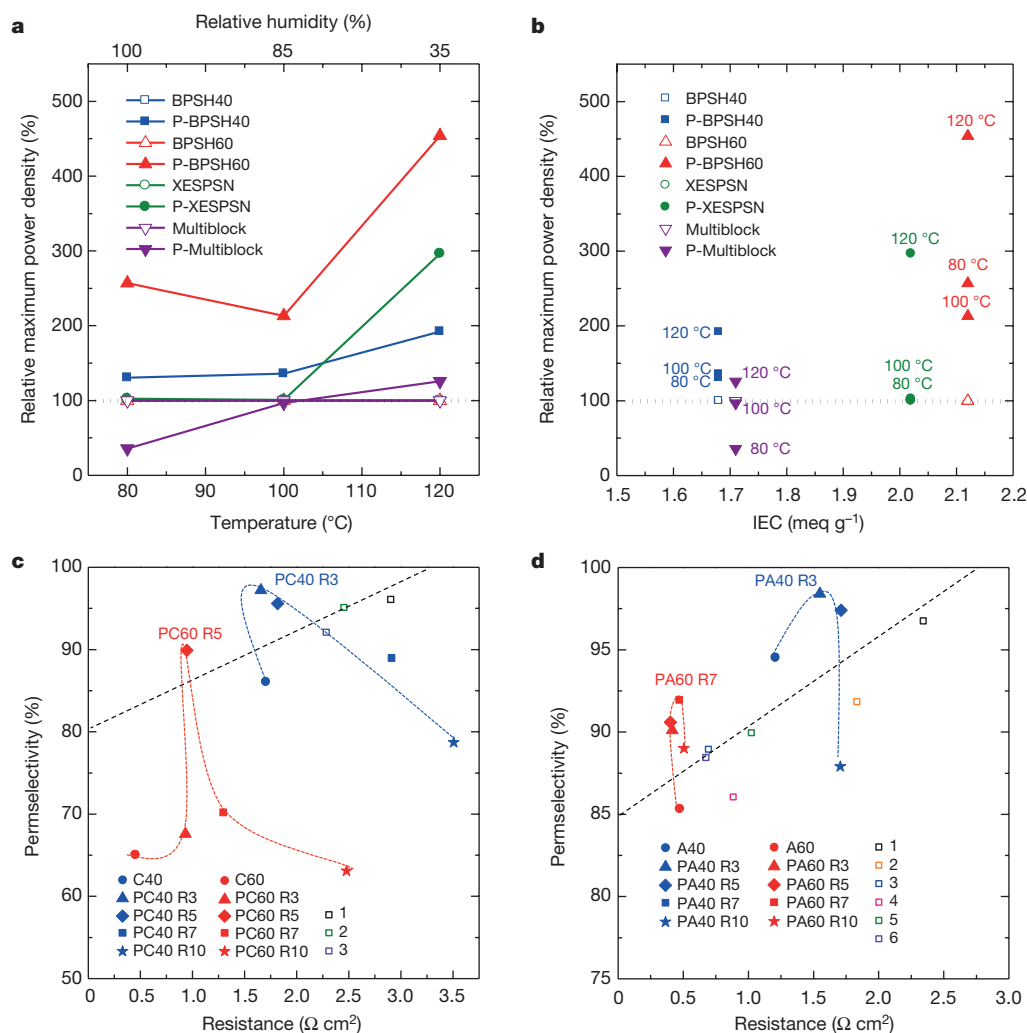


Figure 3 | Effect of the nanovalved surface coating on fuel-cell performance relative to uncoated membranes and selective ion transport behaviours of reverse-electrodialysis membranes. **a**, Single membrane electrode assembly fuel-cell tests were performed at low-to-medium temperature in a low-humidity fuel-cell system. The relative maximum power density ($P_{\text{pc}}/P \times 100$, where P_{pc} and P are the maximum power densities of the plasma-coated membrane and pristine membrane, respectively) increases for various types of aromatic hydrocarbon proton-exchange membranes^{9,23}, as temperature increases from 80 °C to 120 °C with reduced RH from 100% to 35%, under 1.5 atm pressure. **b**, The data are also presented according to the ion-exchange capacity (IEC) values of proton-exchange membranes from 1.68 meq g⁻¹ to 2.12 meq g⁻¹, where 'meq' is milliequivalent, a unit measuring the relative amount of charged functional groups in the polymer matrix. **c**, The permselectivity and membrane resistance of plasma-coated cation-exchange polybiphenylsulfone copolymer (CBPS40 (PC40)) in sodium form and plasma-coated CBPS60 (PC60) (which are sodium-type cation-exchange membranes) with 0.1–0.5 M and 0.5 M NaCl solution for

proton-exchange membranes (Extended Data Fig. 2c), we compared the single fuel-cell performances of uncoated and plasma-coated membranes in Fig. 3a, b (see also Supplementary Discussion 2.7). The efficacy of our conceptual approach is evident at higher temperatures and low-humidity. Uncoated BPSH40 and BPSH60 both show drastic declines in performance with increased temperatures of 100 °C and 120 °C, owing to low water retention, despite their thermal stability (a high glass-transition temperature of over 273 °C; refs 21 and 22). However, the surface-coated P-BPSH40 and P-BPSH60 membranes show a two- and fourfold enhancement in current-voltage performance compared with uncoated membranes (see Fig. 3a, and Extended Data Figs 7b and c, and 8b and c). Apart from water

the reverse-electrodialysis system are presented as curves, left to right, showing the effect of increasing the number of coating cycles from 0 to 10 cycles (R3, R5, R7 and R10). The reference cation-exchange membranes are CMX (from Neosepta/ASTOM; 1, black square), FGD (from Fumacep/Fumatech; 2, green square), CMV (from Selemon/AGC Asahi Glass; 3, purple square). **d**, The permselectivity and membrane resistance of plasma-coated ABPS40 (PA40) and plasma-coated ABPS60 (PA60) (which are chloride-type anion-exchange membranes) are presented by increased number of coating cycles from 0 to 10 cycles (R3, R5, R7 and R10). The reference anion-exchange membranes are AMX (from Neosepta/ASTOM; 1, black square), AM-1 (from Neosepta/ASTOM; 2, orange square), AFN (from Neosepta/ASTOM; 3, blue square), FAD (from Fumacep/Fumatech; 4, red square), DSV (from Selemon/AGC Asahi Glass; 5, green square), APS (from Selemon/AGC Asahi Glass; 6, purple square). Resistances and permselectivity of plasma-coated BPS (P-CBPS, P-ABPS) surpass the trade-off relationship for commercial ion-exchange membranes (black dashed upper-bound line).

retention, the better performance of P-BPSH can be ascribed to improved compatibility between the hydrophobic membrane surfaces and hydrophobic catalyst layers in the membrane electrode assembly. These synergistic effects enable the hydrophobic surface-coated membranes to perform far better than might be expected from the proton conductivity enhancement shown in Extended Data Fig. 9a and b and Supplementary Discussion 2.8. However, the compatibility issue is not strongly dependent on humidity. Accordingly, it is clear that the water-conserving effect is dominant in improving fuel-cell performance under low-humidity conditions. In addition, as we anticipated, this plasma treatment can be applied to various other types of hydrocarbon polymer membranes (Extended Data Fig. 2c) that have water

retention and proton conductivity superior to that of BPSH, such as end-group cross-linked sulfonated random copolymers (XESPSN) and multiblock BPSH-poly(arylene ether sulfone) (BPS) copolymers (Multiblock), leading to further gains in fuel-cell performance (see also Supplementary Discussion 2.7 and Extended Data Fig. 7a–c).

After simple hydrophobic atmospheric plasma treatment for coating the surface of membranes (less than fifteen minutes experimentally), P-BPSH showed large differences in water desorption properties compared with uncoated BPSH. In the nanovalue effect, water retention is achieved by the coating's surface nanocrack pattern, which regulates proton transport and water desorption (Supplementary Discussion 2.9). For hydrocarbon proton-exchange membranes, this nanovalue regulation allows electrochemical fuel-cell performance to be largely maintained under dehydrating conditions of reduced humidity and elevated temperature, particularly over 100 °C (see Extended Data Fig. 9). The intermediate operating temperature of over 100 °C can also maximize the advantageous high glass-transition temperature of hydrocarbon proton-exchange membranes for enhanced membrane stability (Extended Data Fig. 10). This unexpected result is noteworthy because the hydrophobic plasma-coating technique is attractive for commercial scale-up. It can be conducted at conditions of atmospheric pressure, which provides a generally applicable process at room temperature and atmospheric humidity. Although hydrocarbon membranes offer some advantages over perfluorosulfonic acid membranes, they also come with drawbacks such as long-term stability at low humidity, so further investigation is needed to determine whether nanocrack coatings will prove useful in commercial fuel-cell applications.

In another application, the nanocrack concept may be applied to introduce selective ion transport barriers in membranes requiring property trade-off behaviour in an aqueous environment (that is, fully hydrated). Plasma-treated cation-exchange and anion-exchange membranes for reverse electrodialysis show the proof of concept for attaining contradictory properties: high ion selectivity (that is, the surface barrier property) as well as high ion conductivity (that is, low bulk resistance). Membrane performance exceeds that of commercial membranes, as shown in Fig. 3c (for cation-exchange membranes) and Fig. 3d (for anion-exchange membranes) (see also Supplementary Discussion 2.10). Augmented ion transport across the coated cation-exchange and anion-exchange membranes relies both on the optimized layer thickness and on the nanocracked hydrophobic membrane surface, which limits co-ion transport. In particular, the ion-selective coating layer drastically increases the permselectivity of ion-exchange membranes (samples PC60 or PA60; see Fig. 3) while sustaining low membrane resistances, suggesting the potential to surpass the trade-off relationship.

We envisage that the nanovalue effect could be usefully applied to other areas. We have illustrated the nanovalue effect for improving the performance of ion-conducting and reverse-electrodialysis membranes, but we anticipate that the nanocrack effect may provide an elegant solution to improving the performance of polymeric membranes having property trade-off behaviours that differ from the bulk material properties. Examples of where this might apply include controlled drug delivery, ion-exchange membranes (where management of water content in the membrane is critical for controlling the transport properties), membrane contactors, membrane distillation, and pervaporation (that is, hydrophobic membranes under aqueous conditions).

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 16 March 2015; accepted 2 March 2016.

- Appleby, A. J. & Foulkes, F. R. *Fuel Cell Handbook* (Van Nostrand Reinhold, 1989).
- Park, C. H., Lee, C. H., Guiver, M. D. & Lee, Y. M. Sulfonated hydrocarbon membranes for medium-temperature and low-humidity proton exchange membrane fuel cells (PEMFCs). *Prog. Polym. Sci.* **36**, 1443–1498 (2011).

- Li, Q., Jensen, J. O., Savinell, R. F. & Bjerrum, N. J. High temperature proton exchange membranes based on polybenzimidazoles for fuel cells. *Prog. Polym. Sci.* **34**, 449–477 (2009).
- Steele, B. C. H. & Heinzel, A. Materials for fuel-cell technologies. *Nature* **414**, 345–352 (2001).
- Moghaddam, S. *et al.* An inorganic–organic proton exchange membrane for fuel cells with a controlled nanoscale pore structure. *Nature Nanotechnol.* **5**, 230–236 (2010).
- Service, R. F. Newcomer heats up the race for practical fuel cells. *Science* **303**, 29 (2004).
- Kreuer, K.-D., Paddison, S. J., Spohr, E. & Schuster, M. Transport in proton conductors for fuel-cell applications: simulations, elementary reactions, and phenomenology. *Chem. Rev.* **104**, 4637–4678 (2004).
- Chen, Y. *et al.* Enhancement of anhydrous proton transport by supramolecular nanochannels in comb polymers. *Nature Chem.* **2**, 503–508 (2010).
- Lee, S. Y. *et al.* A capillary water retention effect to improve medium-temperature fuel cell performance. *Electrochem. Commun.* **31**, 120–124 (2013).
- Mauritz, K. A. & Moore, R. B. State of understanding of Nafion. *Chem. Rev.* **104**, 4535–4586 (2004).
- Park, C. H. *et al.* Phase separation and water channel formation in sulfonated block copolyimide. *J. Phys. Chem. B* **114**, 12036–12045 (2010).
- Schmidt-Rohr, K. & Chen, Q. Parallel cylindrical water nanochannels in Nafion fuel-cell membranes. *Nature Mater.* **7**, 75–83 (2008).
- Kreuer, K. D. & Portale, G. A critical revision of the nano-morphology of proton conducting ionomers and polyelectrolytes for fuel cell applications. *Adv. Funct. Mater.* **23**, 5390–5397 (2013).
- Zhang, S. *et al.* A review of platinum-based catalyst layer degradation in proton exchange membrane fuel cells. *J. Power Sources* **194**, 588–600 (2009).
- Borup, R. *et al.* Scientific aspects of polymer electrolyte fuel cell durability and degradation. *Chem. Rev.* **107**, 3904–3951 (2007).
- Denes, F. S. & Manolache, S. Macromolecular plasma-chemistry: an emerging field of polymer science. *Prog. Polym. Sci.* **29**, 815–885 (2004).
- Tendero, C., Tixier, C., Tristant, P., Desmaison, J. & Leprince, P. Atmospheric pressure plasmas. *Spectrochim. Acta B* **61**, 2–30 (2006).
- Bonnar, M. P. *et al.* Hydrophobic coatings from plasma polymerized vinyltrimethylsilane. *Chem. Vap. Deposition* **5**, 117–125 (1999).
- Hickner, M. A., Ghassemi, H., Kim, Y. S., Einsla, B. R. & McGrath, J. E. Alternative polymer systems for proton exchange membranes (PEMs). *Chem. Rev.* **104**, 4587–4612 (2004).
- Huang, Y.-H. *et al.* Investigation of fine-structure of polyamide thin-film composite membrane under swelling effect by positron annihilation lifetime spectroscopy and molecular dynamics simulation. *J. Membr. Sci.* **417–418**, 201–209 (2012).
- Kim, Y. S. *et al.* State of water in disulfonated poly(arylene ether sulfone) copolymers and a perfluorosulfonic acid copolymer (Nafion) and its effect on physical and electrochemical properties. *Macromolecules* **36**, 6281–6285 (2003).
- Fan, Y. *et al.* The effect of block length upon structure, physical properties and transport within a series of sulfonated poly(arylene ether sulfone)s. *J. Membr. Sci.* **430**, 106–112 (2013).
- Lee, S. Y. *et al.* Morphological transformation during cross-linking of a highly sulfonated poly(phenylene sulfide nitrile) random copolymer. *Energ. Environ. Sci.* **5**, 9795–9802 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements This research was supported by the Nano-Material Technology Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012M3A7B4049745). C.M.D. is supported by the Australian Research Council (DE40101359). C.M.D., A.W.T. and A.J.H. acknowledge the CSIRO Julius Career award, the CSIRO Office of the Chief Executive Science Leader Scheme and the Australia–Korea Foundation Early Career Researchers Program. M.D.G. is a BK21-Plus visiting professor at Hanyang University.

Author Contributions Y.M.L. conceived the study. C.H.P. and Y.M.L. designed the experiments and C.H.P., S.Y.L., D.S.H., Y.M.L., M.D.G., Tae-Woo K. and Tae-Wuk K. wrote the manuscript. C.H.P., D.S.H. and D.H.C. conducted plasma treatment experiments, X-ray photo electron spectroscopy analysis, and set the coating condition. C.H.P. and S.Y.L. conducted contact-angle measurements and scanning electron microscopy image collecting. D.W.S. conducted dynamic vapour sorption analysis and AFM image collecting. S.Y.L. and K.H.L. conducted electrochemical fuel-cell performances. C.M.D. and A.J.H. conducted PALS analysis. A.W.T. conducted mathematical modelling of water sorption through membranes. Tae-Woo K. and Tae-Wuk K. conducted microscopic observation of cactus stems. M.L. and D.-S.K. conducted the mathematical analysis of surface patterns using the Voronoi diagram program. D.S.H., D.H.C. and K.H.L. fabricated ion-exchange membranes and evaluated the electrochemical performance of the ion-exchange membrane for reverse electrodialysis. C.H.P., S.Y.L., D.S.H., D.W.S., D.H.C., M.D.G. and Y.M.L. discussed the results. All authors commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Y.M.L. (ymllee@hanyang.ac.kr) or M.D.G. (guiver@tju.edu.cn).

METHODS

Materials. 4,4'-Dichlorodiphenylsulfone (DCDPS), 4,4'-difluorodiphenylsulfone (DFDPS), 4,4'-dihydroxybiphenyl (BP) and tetramethyl bisphenol A (TMBP) (Sigma-Aldrich) were purified by recrystallization with ethanol, and dried at 60 °C for one day under vacuum. 3,3'-Disulfonate-4,4'-dichlorodiphenylsulfone (SDCDPS) was prepared via direct sulfonation with fuming sulfuric acid. Fuming sulfuric acid, toluene, dimethylacetamide (DMAc), *N*-methylpyrrolidone (NMP), potassium carbonate (K_2CO_3), methanol (MeOH), 1,1,2,2-tetrachloroethane, toluene and acrylic acid (Sigma-Aldrich) were used as received. *N*-bromosuccinimide (NBS) and benzoyl peroxide (BPO) (Sigma-Aldrich) were used as received for bromination of tetramethylated poly(arylene ether sulfone) (TMBPS). Sodium chloride (NaCl) (Daejung Hwageum Chemical) was used as received.

Synthesis of sulfonated poly(arylene ether sulfone) random copolymer ion-exchange polymers. Sulfonated poly(arylene ether sulfone) (BPS) with 40% and 60% degrees of sulfonation was synthesized via polycondensation polymerization of SDCDPS (4 mmol, 3.93 g for BPSH40; 6 mmol, 5.90 g for BPSH60), DCDPS (6 mmol, 3.45 g for BPSH40; 4 mmol, 2.3 g for BPSH60), K_2CO_3 (11.5 mmol, 3.1787 g), BP (10 mmol, 3.7242 g) in 40 ml DMAc and 20 ml toluene mixture. The mixture was heated at 140 °C and refluxed from 140 °C to 155 °C for 3 h to remove water as a byproduct in the form of an azeotropic mixture with toluene. The reaction was maintained at 165 °C for 20 h and the solution was precipitated in cold deionized water. The precipitate was washed several times with deionized water, boiled in deionized water to remove K_2CO_3 , and then dried at 120 °C for 12 h under vacuum.

Membrane fabrication. BPS membranes in sodium form (BPSNa) were prepared by casting solutions of ~15 wt% polymer in NMP, followed by evaporation under ambient conditions at 40 °C for 12 h, 60 °C for 2 h, 100 °C for 2 h and 120 °C *in vacuo* for 2 h. The subsequent drying protocol was in a vacuum oven sequentially at 45 °C for 24 h, 60 °C for 2 h and 120 °C for 6 h. The resulting BPSNa membrane was transformed to cationic-exchange membrane in sodium form (CBPS; C40 and C60) for reverse electrodialysis by immersing it in 1 M NaCl. The resulting BPSNa membranes were also immersed in boiling deionized water for 2 h to remove the residual solvent, and then treated in boiling 1 M sulfuric acid for 2 h. After washing in boiling deionized water for 4 h, BPS membranes in the protonated form (BPSH) were dried at 120 °C for 12 h under vacuum. The approximate thickness of the membranes was 50 μ m.

Fabrication of anion-exchange membrane. Tetramethylated poly(arylene ether sulfone) (TMBPS) was synthesized via condensation polymerization of DFDPS (30 mmol, 7.6275 g), TMBP (12 mmol, 2.9077 g) for tetrabrominated poly(arylene ether sulfone) with degree of functionalization 40 (TBrBPS40) and 18 mmol, 4.3616 g for TBrBPS60, BP (18 mmol, 3.3518 g for TBrBPS40; 12 mmol, 2.2345 g for TBrBPS60) and K_2CO_3 (45 mmol, 6.2195 g) in 85 ml DMAc and 80 ml toluene mixture. The mixture was heated at 120 °C and refluxed for 4 h to remove water with toluene via the azeotropic method. The reaction temperature was increased to 160 °C and maintained for 2 h. After cooling to room temperature, the mixture was precipitated in cold deionized water. The resulting TMBPS precipitate was washed several times with deionized water and dried at 60 °C *in vacuo*. TMBPS polymer (14.1300 g), NBS (60.14 mmol, 7.1674 g for aminated poly(arylene ether sulfone) copolymer with degree of functionalization 40 (ABPS40); 87.8837 mmol, 10.4723 g for ABPS60) and BPO (4.010 mmol, 1.2950 g for ABPS40; 5.8589 mmol, 1.8923 g for ABPS60) were introduced in 127 ml of 1,1,2,2-tetrachloroethane under an argon atmosphere for the brominated TMBPS (TBrBPS). The mixture was heated to 80 °C and was maintained at that temperature for 12 h. After cooling to room temperature, the mixture was precipitated with methanol and washed several times. The resulting TBrBPS polymer was dried and dissolved in 15 wt% NMP and the solution was cast onto a glass plate. The polymer solution was evaporated slowly in an oven as the temperature was increased to 40 °C for 24 h, to 60 °C for 2 h, and to 120 °C for 2 h *in vacuo*. The resulting membrane was immersed in trimethylamine 45% (w/w) aqueous solution for 24 h to substitute bromide with tetramethyl ammonium functional groups, which allow conduction of hydrated chloride ions through the polymer membranes. The resulting membrane was transformed to its chloride form (ABPS; A40 and A60) for reverse electrodialysis by immersing it in 1 M NaCl.

Atmospheric plasma treatment. Atmospheric non-equilibrium plasma treatment can be employed to deposit a thin polymeric layer on the membrane via plasma polymerization for hydrophobic surface modification^{16–18}. Electric discharge powered by radio frequency generates various monomer fragment species such as neutral molecules, ions and radicals, which are grafted or adsorbed onto the surface. Grafted oligomers and radicals grow from the polymer surface as the plasma discharge continues via plasma polymerization. Plasma-enhanced chemical vapour deposition imparts various chemical properties originating from different precursors, creating hydrophobicity and ion conductive and physical membrane properties compared to non-treated membranes.

In this work, an atmospheric plasma discharge system (SHP-1000, APP) with a 13.56-MHz radio frequency discharger was used for hydrophobic plasma surface coating. Dried BPSH membrane film (10 cm \times 10 cm) was attached onto an aluminium plate. The aluminium plate traverse speed was controlled to 30 mm s⁻¹ along the y axis under the plasma glow discharging source (18 cm \times 2 cm) with the gap set at about 2.5 mm, which allowed homogeneously coated surfaces to be obtained. Atmospheric plasma treatment (input power of 150 W) was performed under controlled chamber conditions of atmospheric pressure at 25 °C and 40% RH, with gas flow rates of 10 ml min⁻¹ of octafluorocyclobutane (*c*-C₄F₈) and 20 litres per minute of He as coating cycles increased from 3 cycles to 40 cycles (Extended Data Fig. 2a). For graft plasma polymerization, the radio-frequency discharger excites the He carrier gas, which lowers the surface energy of the BPSH, CBPS or ABPS membrane (Extended Data Fig. 2c) and *c*-C₄F₈ gas molecules are also excited and fragment into various types of species including neutral molecules, radicals and ions of C_xF_y in the plasma field (Extended Data Fig. 2b). The generated monomer fragments undergo plasma polymerization, including radical polymerization, on the excited surface of BPSH or CBPS and ABPS, which continues to grow fluorocarbon graft branches with additional coating cycles (Extended Data Fig. 2b). The thickness of the deposited polymer layer as well as the hydrophobicity of the membrane surface is controlled by the number of coating cycles.

X-ray photoelectron spectroscopy. X-ray photoelectron spectroscopy was acquired using a Sigma Probe (Thermo VG Scientific) equipped with an Al K α monochromatic X-ray source under a base chamber pressure of 5×10^{-8} mbar. The spectrum for each atom (such as C, S, O and F) was fitted using the Spectral Data Processor Version 7 program (<http://xpsdata.com>) in order to estimate the atomic composition change on the membrane surface. The spectra for sulfur and fluorine atoms were scanned from 155 eV to 185 eV and from 679 eV to 699 eV, respectively, in stepwise increases of 0.10 eV.

Atomic force microscopy. AFM images were obtained using a Digital Instruments MultiMode 8 AFM (Veeco) with a NanoScope V controller (Veeco). A silicon probe (Nanosensors) with a force contact of 1.2–29 N m⁻¹ was used to scan surface morphology. Samples were conditioned at different hydration conditions to present the surface morphology changes of the membranes. Plasma-coated samples were dried immediately after treatment at 120 °C *in vacuo* for 12 h before AFM measurement. Assist scan tapping mode was performed in a fluid cell filled with deionized water in order to scan the surface of fully hydrated plasma-coated membranes, which were immersed in deionized water for 24 h at room temperature. Hydrated samples were allowed to air-dry on the piezo-scanner of the microscope for at least 30 min to obtain the surface morphology of partially dehydrated membranes. AFM was performed in a fluid cell filled with deionized water using the assist scan tapping mode and under atmospheric conditions with 30% to ~45% RH using the tapping mode. Surface topological depths were estimated from AFM height images.

Dynamic vapour sorption. DVS was performed using a DVS Advantage (Surface Measurement Systems, UK) with Cahn balance D-200 to investigate water vapour sorption and desorption behaviour. The RH was controlled by changing the ratio of dry and wet gases in the range of 0%–90%. All membrane samples were dried at 120 °C under vacuum for 12 h to remove absorbed water before measurements.

Contact-angle measurement. Water contact angles were measured on an Easy Drop Standard (Krüss) instrument equipped with a charge-coupled device camera by the sessile-drop method. The average volume of water droplet was about 2 μ l for contact angle measurements.

Scanning electron microscopy. Field-emission scanning electron microscopy (JSM-7900F, JEOL, Japan) was performed to investigate the morphology of the hydrophobic coated surface layer.

Positron annihilation lifetime spectroscopy. PALS is a nuclear technique used to measure the free volume of bulk polymer materials. The lifetime of the *ortho*-positronium (*o*-Ps, the bound state of a positron and an electron of the same spin) is related to the size of the free volume elements within the polymer. The *o*-Ps will preferentially locate within the pore spaces and will then annihilate with an electron from the pore wall of the sample. The size of the pore determines how long it takes for an annihilation event; larger pores result in longer lifetimes. The Tao–Eldrup equation is a quantum model used to calculate the average spherical pore size from the *o*-Ps lifetime²⁰. The lifetime has an associated intensity value which is related to the relative number of pores within the sample. The samples were stacked 2 mm thick each side of a positron point source (30 μ Ci of ²²NaCl enclosed in a Mylar envelope). The volume of plasma surface coating in the 2 mm film stack (for a 100 nm coating) is 0.4%, to ensure that the *o*-Ps is probing the bulk membrane. The samples were run under a nitrogen atmosphere with humidity control. The spectra were detected using an automated ORTEC EG&G instrument (Oak Ridge) and analysed using LT version 9 software²⁴. The first lifetime was fixed to 0.125 ns owing to *para*-positronium annihilation (a bound state of an electron

and a positron of opposite spin) and the second lifetime was ~ 0.4 ns owing to free annihilation with an electron. The third component, *o*-Ps, was used to calculate the lifetime, which is related to pore size, and intensity, which is related to the relative number of pores within membranes.

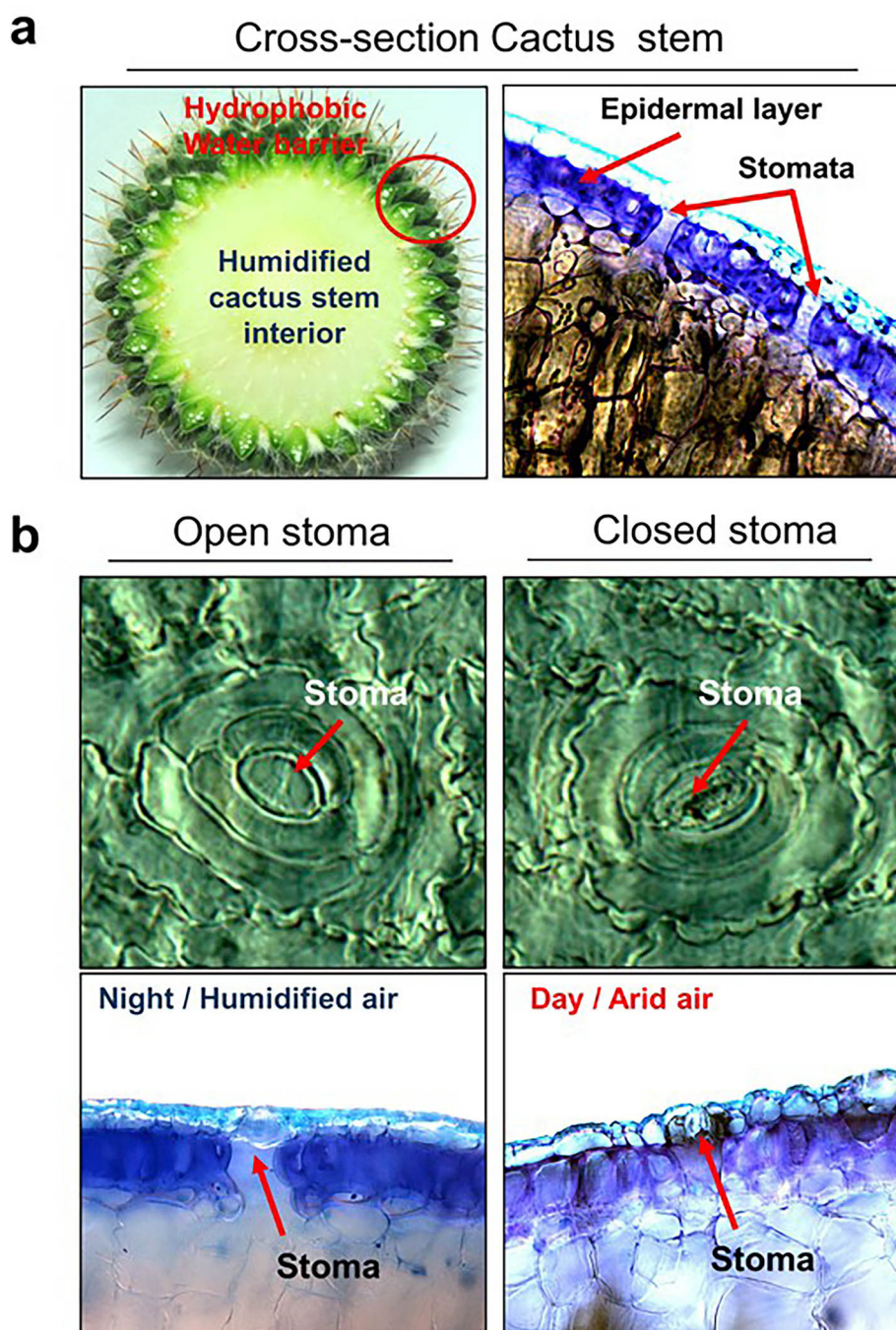
The results show that there is no change in the bulk properties of the membranes caused by the surface coating. The intensity consistently drops with increasing RH owing to the uptake of water and hence a decrease in the number of empty pores. The lifetimes show a more complex trend. The initial drop (0%–30% RH) is probably due to a pore-filling effect and then the increase is due to swelling above 40% RH.

Fuel-cell polarization measurements. The membrane electrode assembly with an active area of 5 cm^2 and Pt loading of 0.5 mg cm^{-2} was fabricated via the screen printing method²⁵. The membrane electrode assembly was set into a single cell test fixture and mounted in a commercial fuel-cell testing station (SMART PEMFC test system, WonATech) which was supplied with temperature- and humidity-controlled gases. The test system was operated at 80°C under various RH conditions with hydrogen and oxygen/air gases under ambient pressure. Additionally, membrane electrode assembly electrochemical performance was evaluated at 80°C (100% RH), 100°C (85% RH) and 120°C (35% RH) under 1.5 atm pressure. Hydrogen gas and oxygen/air gas were supplied at a flow rate considering the stoichiometry ratio on the increased current load.

Fuel-cell durability measurement. A single cell stability test was performed by measuring the variation of current density at constant 0.7 V as a function of time at 120°C and 35% RH. The stability tests were performed three times for each type of membrane. Other conditions were as described above.

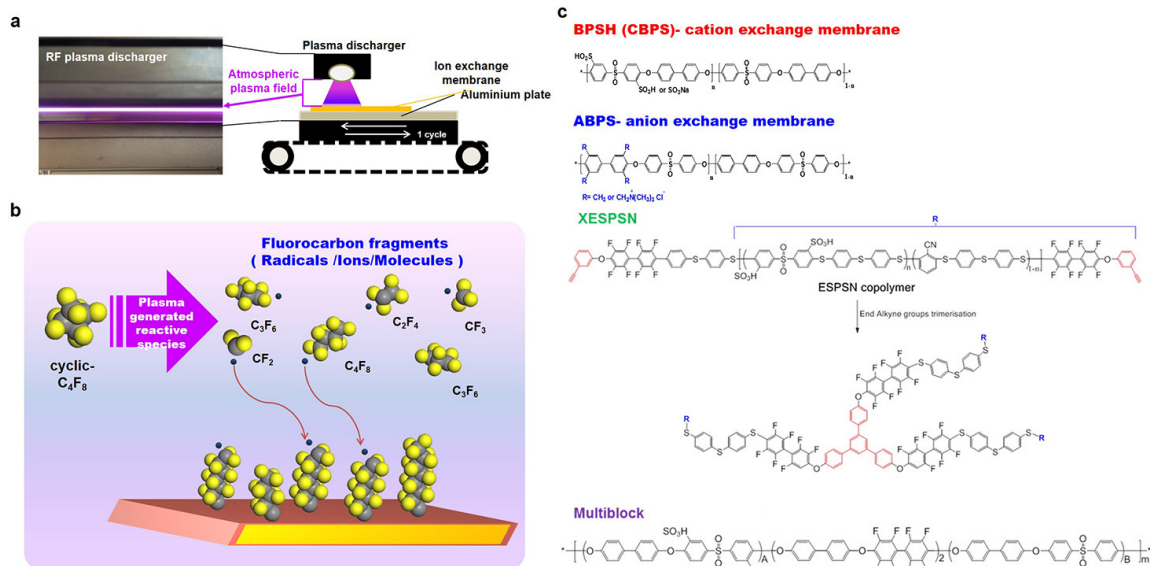
Reverse-electrodialysis measurements. The plasma-coated membranes and uncoated membranes in sodium chloride salt form were soaked in distilled water for at least 24 h to allow complete hydration before measuring permselectivity and membrane resistances in aqueous NaCl solution (0.1 M and 0.5 M) at 20°C . Permselectivity was measured by a static potential method in a two-compartment cell separated by the membrane, where each cell contained 0.1 M and 0.5 M NaCl solution, respectively. Membrane resistances were tested by using an electrochemical impedance spectroscopy analyser in a two-compartment cell equipped with Ag/AgCl reference electrodes (RE-1B, ALS) in 0.5 M aqueous NaCl solution at 20°C .

24. Kansy, J. *et al.* Microcomputer program for analysis of positron annihilation lifetime spectra. *Nucl. Instrum. Methods Phys. Res. Sect. A* **374**, 235–244 (1996).
25. Hwang, D. S., Park, C. H., Yi, S. C. & Lee, Y. M. Optimal catalyst layer structure of polymer electrolyte membrane fuel cell. *Int. J. Hydrogen Energy* **36**, 9876–9885 (2011).
26. Kim, J.-K. *et al.* Voronoi diagrams, quasi-triangulations, and beta-complexes for disks in R2: the theory and implementation in BetaConcept. *J. Comput. Design Eng.* **1**, 79–87 (2014).



Extended Data Figure 1 | Cactus (*Ferocactus schwarzii*) stomata control mechanism for water retention, analogous to self-controlled nanovalve mechanism of plasma-treated membranes. a, Transverse-section microscopic image of cactus stem ($\times 400$ magnification, Normarski differential interference contrast microscopy (Nikon, Eclipse 80i)) illustrates the structure of outer photosynthetic tissues which consist of the impermeable epidermal layer and stomata. **b,** Microscopic pictures

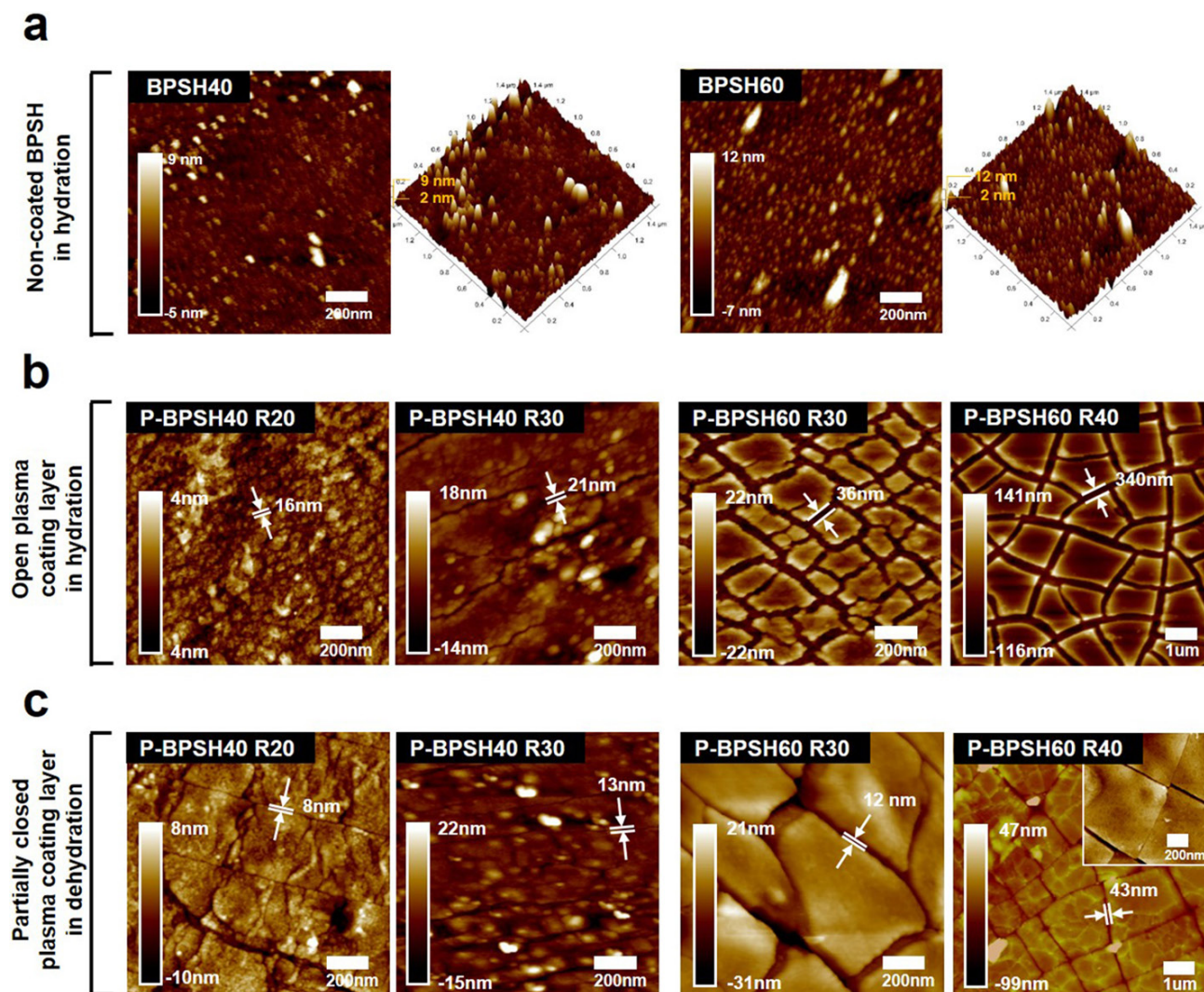
illustrating the cactus water-conserving effect based on the stoma self-control mechanism by swelling of stoma cells. To obtain the transverse section images of the cactus stem, cross-sectional bulky pieces of cactus stem were fixed with fixation solution. After washing with 20% ethanol, thin transversely sliced tissues were cleared with chloral hydrate solution, and this was followed by staining with 0.05% (w/v) toluidine blue O solution for microscopic observation.



Extended Data Figure 2 | Schematic outline of plasma polymerization.

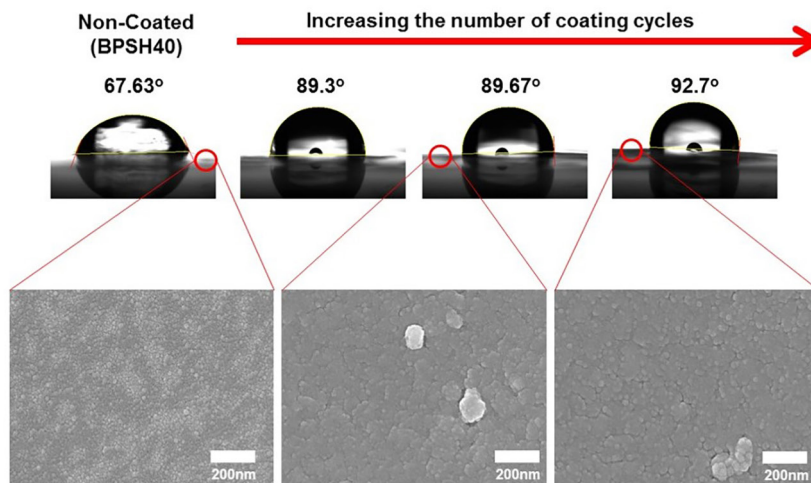
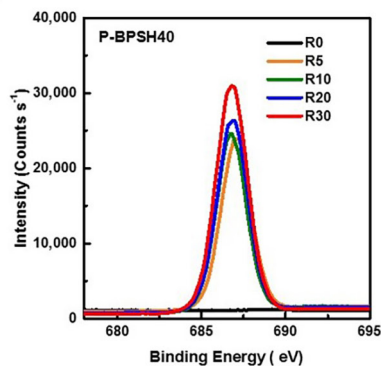
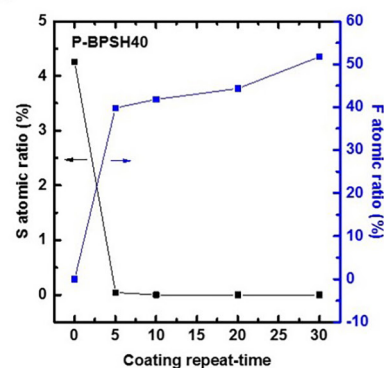
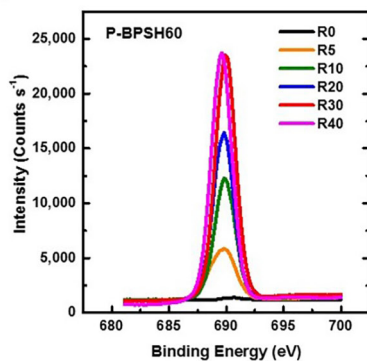
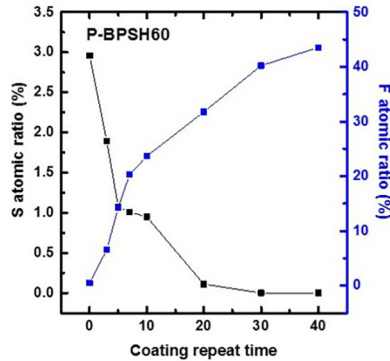
a, Purple-lit atmospheric plasma of *c*-C₄F₈ and He (photograph) was generated from a radio-frequency (RF) glow discharger. **b**, Mechanism of the plasma polymerization and fluorohydrocarbon coating layer. *c*-C₄F₈ was broken down into various species of fluorocarbon monomers. Plasma polymerization formed a hydrophobic coating layer on the nanometre scale. **c**, Polymer structure of hydrocarbon aromatic ion-exchange polymer membranes. Two different degrees of functionalization ($n = 0.4$ or 0.6)

were fabricated in sulfonated poly(arylene ether sulfone) in the protonated form (BPSH) or sodium form (CBPS) and an aminated poly(arylene ether sulfone) (ABPS), respectively. The chemical structures of various types of hydrocarbon aromatic proton-exchange polymer membranes are shown. XESPSN is an end-group cross-linked sulfonated random copolymer with a degree of sulfonation of 0.6. 'Multiblock' refers to a multiblock BPSH-BPS copolymer, which consists of 10 kg mol⁻¹ BPSH and 5 kg mol⁻¹ BPS with a spacer linkage.



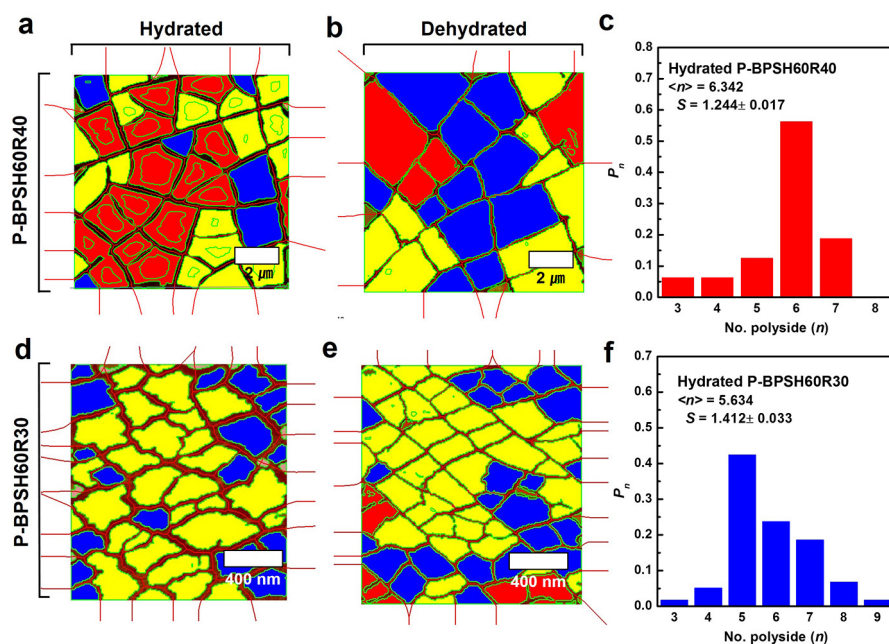
Extended Data Figure 3 | Nanovalved plasma coating layers in response to humidity conditions. a–c, AFM images showed that nanometre-sized cracks were formed upon hydration of plasma-coated membranes with different surface morphology, depending on the degree of sulfonation and coating cycles, when compared to uncoated membranes in hydration (a). The colour scale indicates the topology of surface (depth) according to colour gradation (with the deepest point shown as black and the highest point as bright yellow). R20, R30 or R40 indicates the number of repeated

coating cycles. The water-conserving mechanism with self-control of nanovalves was investigated by the comparison of AFM images between hydrated (b) and dehydrated (c) plasma-coated membranes. The surface morphology of hydrated P-BPSH40 to P-BPSH60 was investigated at different scales from AFM images of dehydrated P-BPSH in order to clarify the entire surface morphology. P-BPSH60 R40 had the highest volumetric swelling ratio, making the morphology change particularly evident.

a**b****c****d****e**

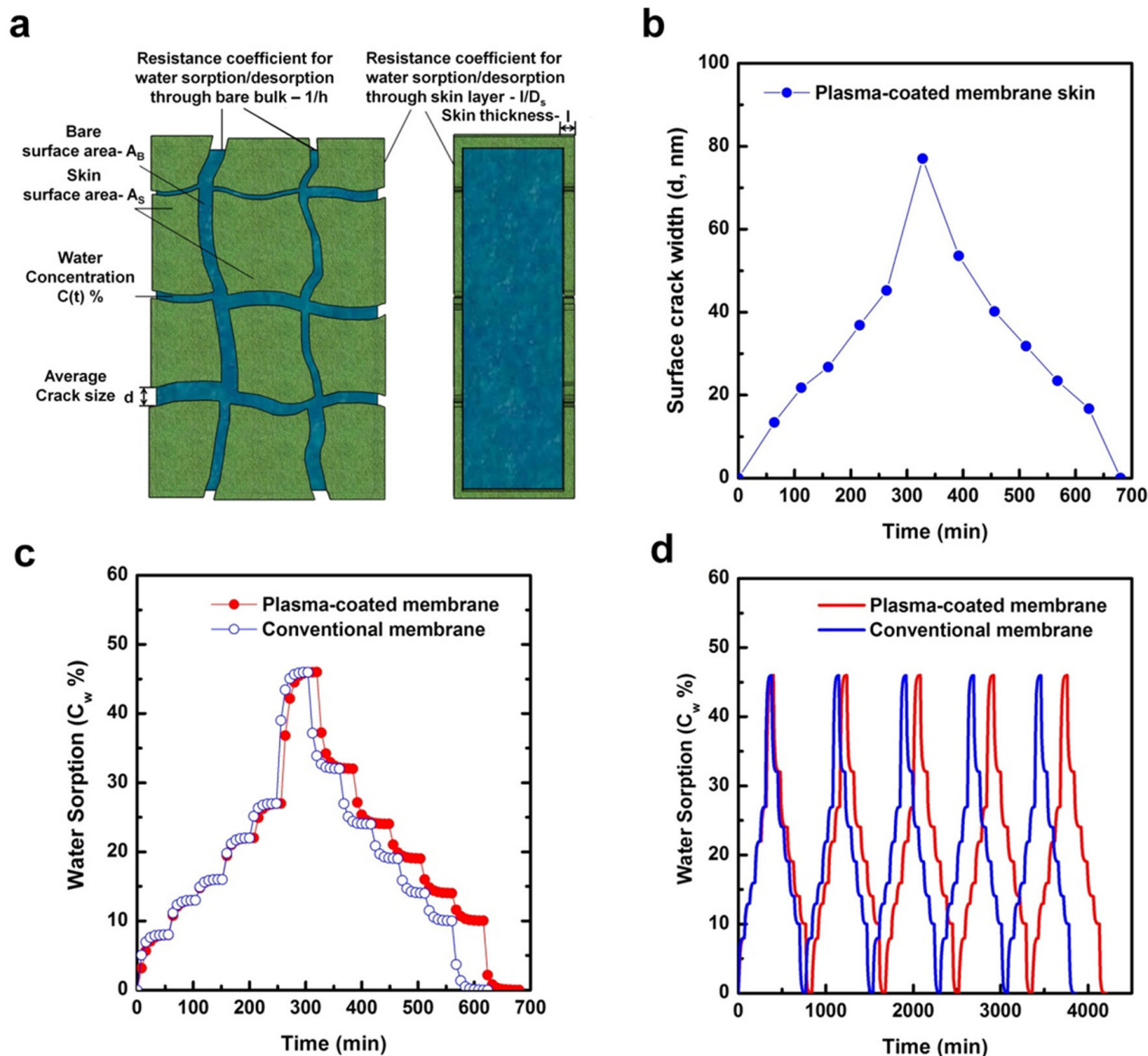
Extended Data Figure 4 | Hydrophobic surface coating layer by plasma treatment. **a**, Contact angles of the hydrophobic atmospheric plasma surface-coated membrane (BPSH40) in specific coating cycles, where the number of coating cycles increases from left to right images, 0, 10, 20, 30 cycles, respectively. **b**, **d**, Spectral change of P-BPSH40 (**b**) and

P-BPSH60 (**d**) in fluorine peak after surface coating measured with X-ray photoelectron spectroscopy. **c**, **e**, Composition change of an F atom included in fluorocarbon layer and an S atom in sulfonic acid groups at surface of P-BPSH40 (**c**) and P-BPSH60 (**e**) with increased coating cycles.



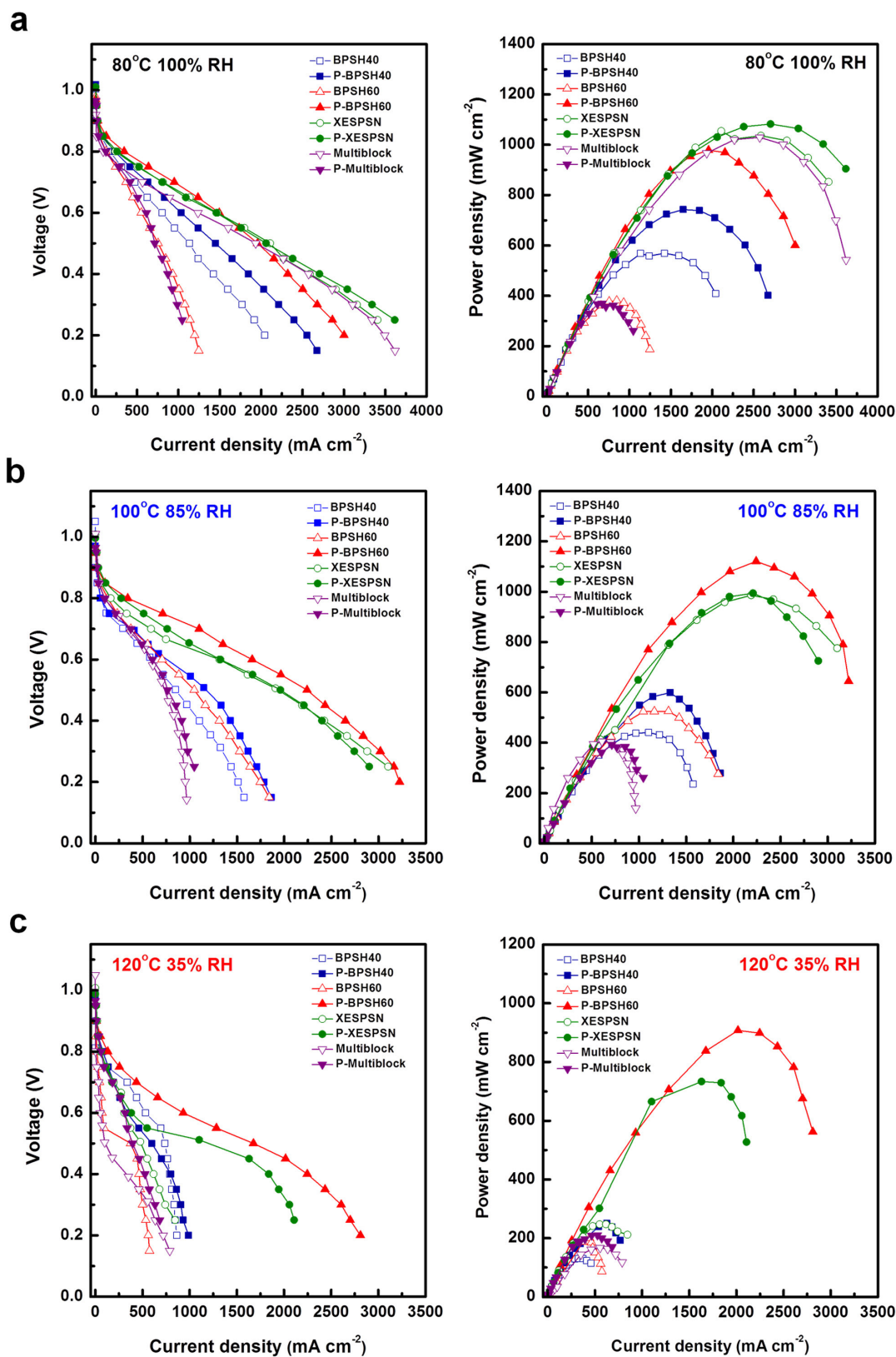
Extended Data Figure 5 | Reproducibility and regularity of nanocrack pattern evaluated via Voronoi tessellation entropy differentiation. AFM images of P-BPSH60R40 (a, b, c) and P-BPSH60R30 (d, e, f) were analysed by area Voronoi diagram. a, b, P-BPSH60R40 hydrated in deionized water (size 10 μm , a), and dehydrated (size 10 μm , b) under 30% to 45% RH conditions. d, e, P-BPSH60R30 hydrated in deionized water (size 1.5 μm , d) and dehydrated (size 1.5 μm , e) under 30% to 45% RH conditions. The circularity of each pattern is indicated in different colours. Blue, yellow, and red colours present high, middle, and low circularity, respectively. The circularity of patterns is defined as the ratio of the circumference of circle in the same area of each pattern to the length of a pattern's boundary. c, f, Tessellation entropy values were evaluated by

probability of frequency distribution from the number of polyside patterns (approximately 20–40) in the AFM images for hydrated P-BPSH60R40 (c) and hydrated P-BPSH40R30 (f). The probability p_n that a Voronoi cell will have n neighbour Voronoi cells, where n is a non-negative integer, is the number of n -polyside patterns divided by the number of total polyside patterns in an image, where n is 3–10. The value of the tessellation entropy was determined by the distribution of probability $S = -\sum p_n \ln p_n$. The average of the number of neighbouring Voronoi cells is $\langle n \rangle = \sum n N_n / \sum N_n$, where N_n is the frequency of the n -polyside pattern. Standard deviations of tessellation entropy were calculated from sixteen samples for each plasma treatment condition by repeating the same plasma coating twice.



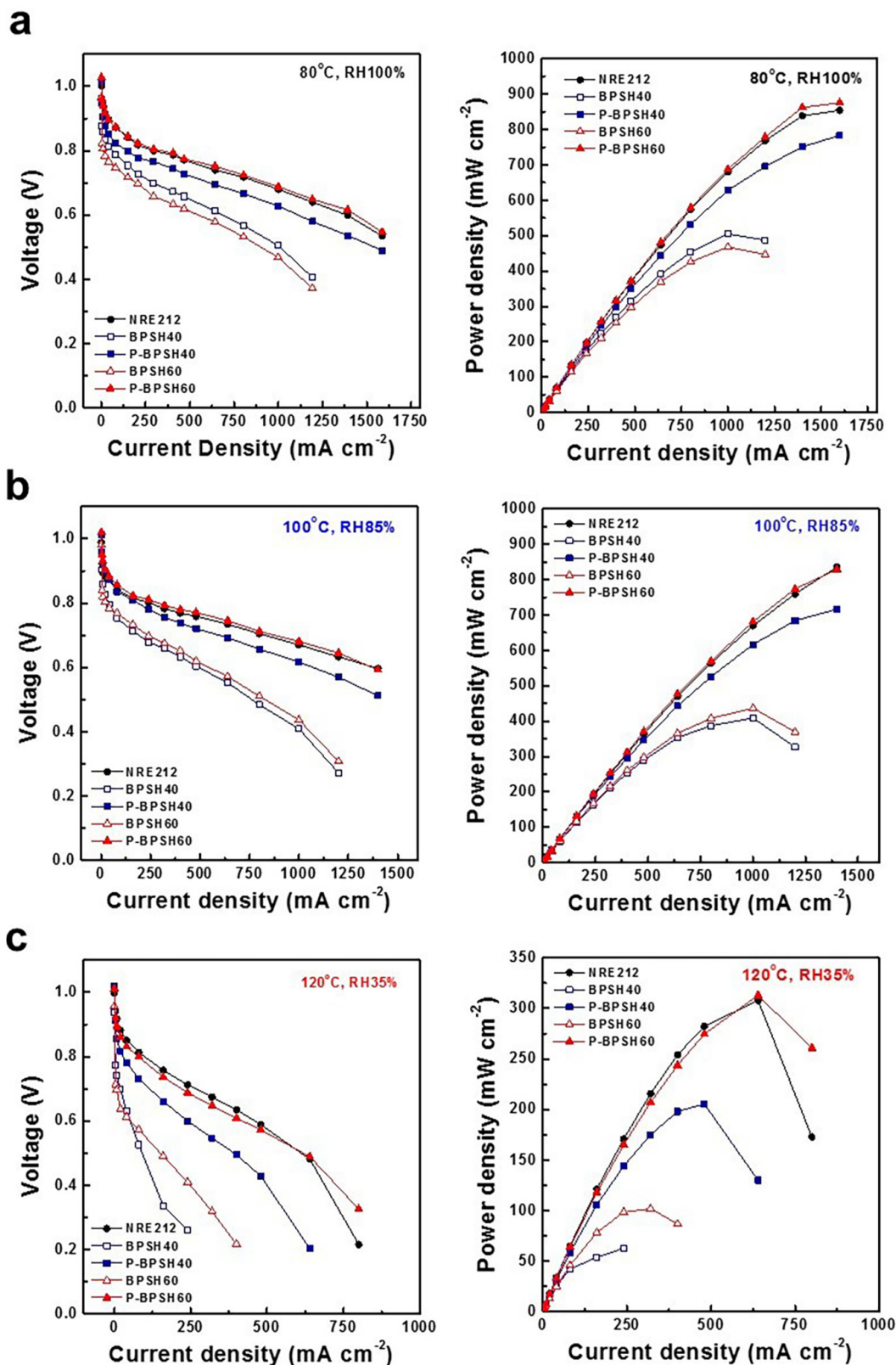
Extended Data Figure 6 | Mathematical model for water sorption-desorption of a plasma-coated membrane. **a**, Variables and configuration for the ordinary differential equation that describes water sorption-desorption for a plasma-coated membrane. **b**, Simulated crack width expansion during DVS time. **c**, **d**, Simulated DVS as a function of time for the conventional membrane and the plasma-coated membrane for a single DVS cycle (**c**) and five pulsatile DVS cycles (**d**). Using equations

(4) and (10) in Supplementary Discussion 2.5, with experimentally determined equilibrium water concentration C_w values, normalized $A = 1$, $h = 1$, $D_s = 0.005$, skin thickness $l = 1$ and step tolerance 0.01242. The hydrophobic skin of the plasma-coated membrane delays water sorption and more importantly inhibits the desorption of water such that the membrane remains hydrated for a longer time compared with the uncoated conventional membrane.



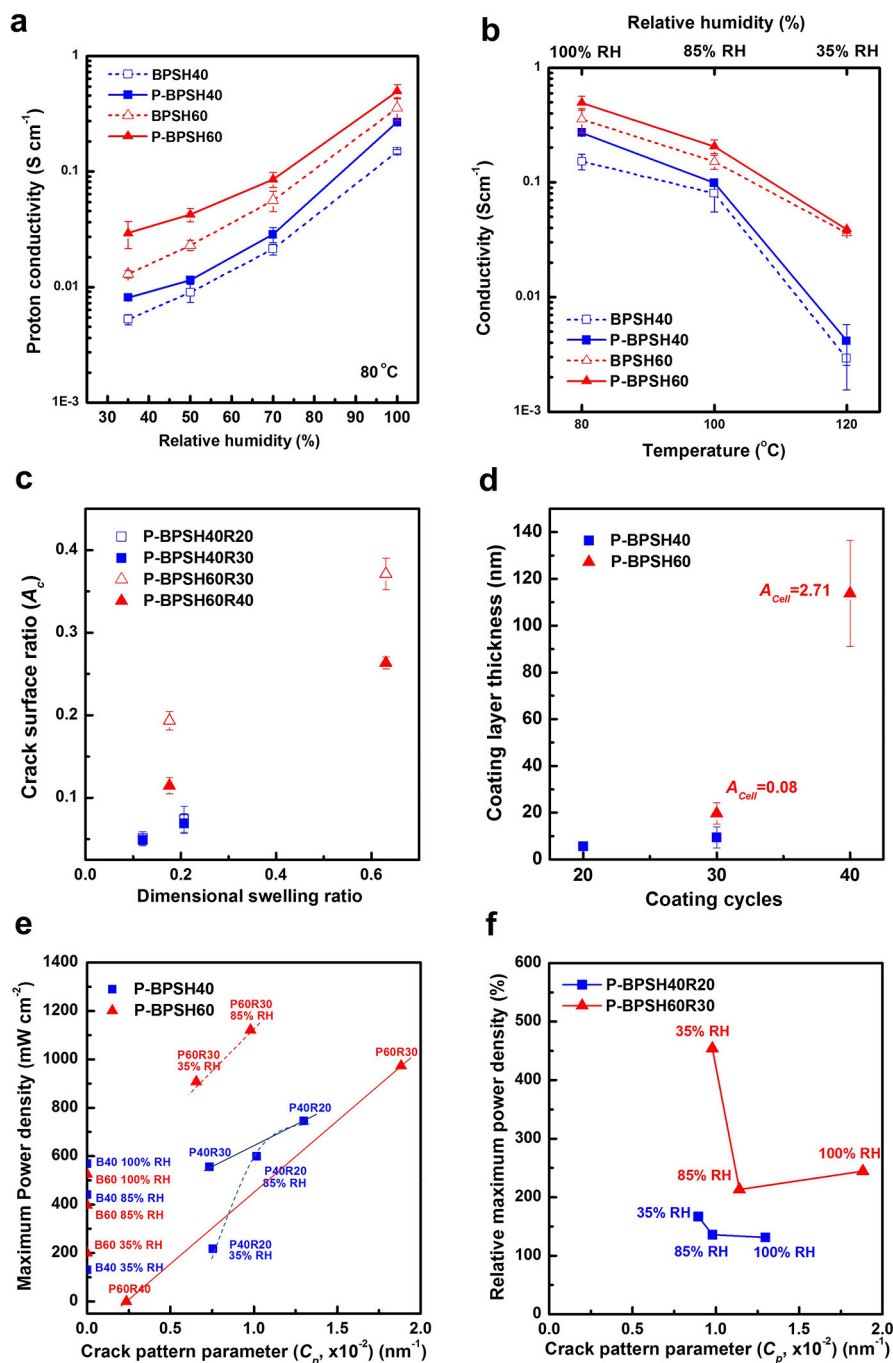
Extended Data Figure 7 | Current–voltage polarization curves for various types of hydrocarbon aromatic proton-exchange membranes in single membrane electrode assembly tests. Fuel-cell performances of uncoated membranes (BPSH40, BPSH60, XESPSN, Multiblock) and plasma-coated membranes (P-BPSH40, P-BPSH60, P-XESPSN,

P-Multiblock) were measured at various RH values and operating temperatures. **a**, At 80 °C under 100% RH and 1.5 atm of pressure. **b**, At 100 °C under 85% RH and 1.5 atm of pressure. **c**, At 120 °C under 35% RH and 1.5 atm of pressure by supplying H₂ and O₂.



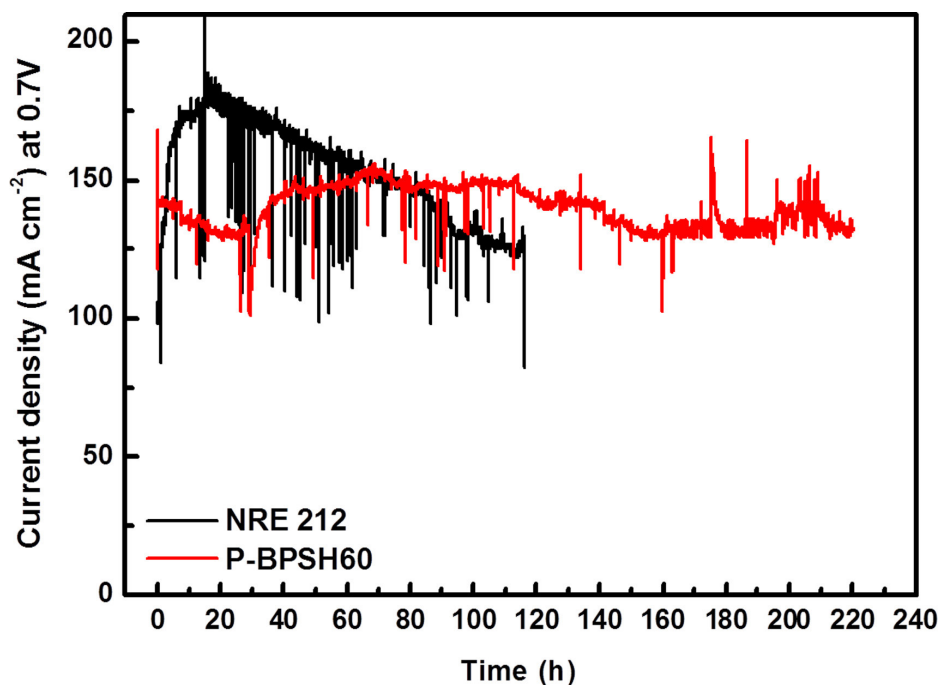
Extended Data Figure 8 | Current-voltage polarization curves of various types of hydrocarbon aromatic proton-exchange membranes in single membrane electrode assembly tests. Fuel-cell performances of uncoated membranes (BPSH40, BPSH60, Nafion NRE212) and plasma-coated membranes (P-BPSH40, P-BPSH60) were measured at various RH

values and operating temperatures. **a**, At 80 °C under 100% RH and 1.5 atm of pressure. **b**, At 100 °C under 85% RH and 1.5 atm of pressure. **c**, At 120 °C under 35% RH and 1.5 atm of pressure by supplying an H₂/air feed.



Extended Data Figure 9 | Effect of nanocrack pattern on electrochemical performances along with proton conductivities and water retention of plasma-coated membranes. Proton conductivity (4-probe method, in-plane) of uncoated membranes (BPSH40, BPSH60) and plasma-coated membranes (P-BPSH40, P-BPSH60) were measured at various RH values and temperatures. **a**, Proton conductivities were measured at 80°C as RH decreased from 100%, 70% and 50% to 33%. **b**, Proton conductivities were measured at 80°C (100% RH), 100°C (85% RH) and at 120°C (35% RH) under 1.5 atm pressure. Values in **a** and **b** are averages of at least fifteen replicates; error bars represent 1 s.d. **c**, **d**, A conceptual crack pattern parameter ($C_p = A_c/I$) is correlated to electrochemical performance, and is defined by crack formation underlying parameters of crack surface area ratio (**c**) (A_c , average values of crack area to total membrane area, were calculated from the Image J program version 1.50b and image processing and analysis was done in the open-source program Java; <https://imagej.nih.gov/ij/index.html>) increased by membrane dimensional

swelling ratio and thickness of coating layer (**d**) (I was measured by AFM image analysis). The average area of one Voronoi cell component (coloured domain with green boundary line in Extended Data Fig. 5a, b, d and e), A_{cell} (in μm^2) was calculated by using the open-source BetaConcept program for Voronoi diagram analysis (<http://voronoi.hanyang.ac.kr/software.htm#BetaConcept>)²⁶. Values in **c** and **d** are averages of at least sixteen replicates; error bars represent 1 s.d. **e**, Electrochemical maximum power densities of P-BPSH40 and P-BPSH60 are proportionally enhanced with increasing C_p . The maximum power densities of uncoated BPSH40 (B40) and BPSH60 (B60) at various RH conditions of 100%, 85% and 35% RH are presented on the y axis, with their crack pattern parameters. **f**, The water retention effect by nanovalue control at low RH is reflected by the relative maximum power density of membrane electrode assembly (in Fig. 3a and b) with correlation of crack pattern parameters at various membrane electrode assembly operating conditions (100% RH at 80°C , 85% RH at 100°C , and 35% RH at 120°C).



Extended Data Figure 10 | Long-term stability test results. Membrane electrode assembly single cell operation on conditions: current density measured at 0.7 V (constant voltage) at 120 °C under 35% RH and 1.5 atm of pressure by supplying H₂ and air. Stability tests were performed three times for each type of membrane. Nafion (NRE212) reaches a maximum current density of 180 mA cm⁻² after 20 h and shows a transient decline in current density. After 120 h, a complete loss of performance occurred and the Nafion membrane appeared to be decomposed. P-BPSH60

maintains its current density of about 150 mA cm⁻² until 220 h, when the measurements were intentionally stopped for observation. The time until a 10% loss of current density occurs is determined to be 60 h and 220 h for Nafion and P-BPSH, respectively. After autopsy of each membrane electrode assembly, the Nafion membrane had a black colour indicating degradation, whereas the P-BPSH membrane still maintained its shape and original colour.

The pentadehydro-Diels–Alder reaction

Teng Wang¹, Rajasekhar Reddy Naredla¹, Severin K. Thompson¹ & Thomas R. Hoye¹

In the classic Diels–Alder [4 + 2] cycloaddition reaction¹, the overall degree of unsaturation (or oxidation state) of the 4 π (diene) and 2 π (dienophile) pairs of reactants dictates the oxidation state of the newly formed six-membered carbocycle. For example, in the classic Diels–Alder reaction, butadiene and ethylene combine to produce cyclohexene. More recent developments include variants in which the number of hydrogen atoms in the reactant pair and in the resulting product is reduced² by, for example, four in the tetrahydro-Diels–Alder (TDDA) and by six in the hexadehydro-Diels–Alder (HDDA)^{3–7} reactions. Any oxidation state higher than tetrahydro (that is, lacking more than four hydrogens) leads to the production of a reactive intermediate that is more highly oxidized than benzene. This increases the power of the overall process substantially, because trapping of the reactive intermediate^{8,9} can be used to increase the structural complexity of the final product in a controllable and versatile manner. Here we report an unprecedented overall 4 π + 2 π cycloaddition reaction that generates a different, highly reactive intermediate known as an α ,3-dehydrotoluene. This species is in the same oxidation state as a benzyne. Like benzyne, α ,3-dehydrotoluenes can be captured by various trapping agents to produce structurally diverse products that are

complementary to those arising from the HDDA process. We call this new cycloisomerization process a pentadehydro-Diels–Alder (PDDA) reaction—a nomenclature chosen for chemical taxonomic reasons rather than mechanistic ones. In addition to alkynes, nitriles (RC \equiv N), although non-participants in aza-HDDA reactions, readily function as the 2 π component in PDDA cyclizations to produce, via trapping of the α ,3-(5-aza)dehydrotoluene intermediates, pyridine-containing products.

The overall oxidation states of the π -bond-containing pair of reactants in Diels–Alder processes can be viewed as the total amount of ‘dehydroness’ (see refs 10, 11) of those species (Fig. 1). This can be identified either from the overall hydrogen atom count or by the number of *sp*-hybridized carbon atoms that engage to create the newly formed six-membered ring (compare the carbon atoms represented by black circles and by black triangles in reactants 4 + 5 and 4 + 8 in Fig. 1). It occurred to us that a 6 π -electron net [4 + 2] cycloaddition of reactants containing a total of five *sp*-hybridized carbon atoms—namely, an allenyne + alkyne pair like 11 + 4—might produce an α ,3-dehydrotoluene (see 12, Fig. 1d). The parent species 12 itself has been characterized by photoelectron spectroscopy in the gas phase¹², and derivatives of 12, which comprise a little-explored class of reactive

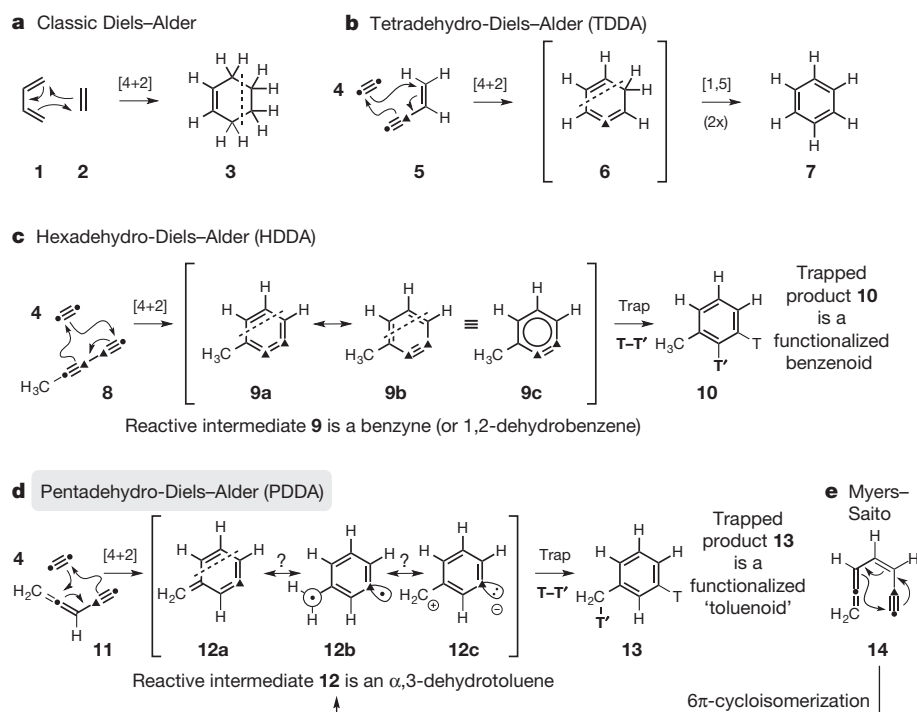


Figure 1 | Terminology associated with various cyclizations in the Diels–Alder family of 4 π + 2 π reactions. a, The classic example of a diene (1,3-butadiene, 1) and a dienophile (ethylene, 2) reacting to give a six-membered cyclic alkene (cyclohexene, 3). **b**, The absence of four hydrogen atoms gives the tetrahydro (TDDA) variant; the product is in the benzene oxidation state. **c**, The absence of six hydrogen atoms gives

the hexadehydro Diels–Alder (HDDA) variant. **d**, The unprecedented pentadehydro-Diels–Alder (PDDA) reaction proceeds via an α ,3-dehydrotoluene (see 12); importantly, both the HDDA and PDDA reactions result in formation of trappable reactive intermediates. **e**, α ,3-Dehydrotoluenes have previously been generated principally by cyclization of allenyl enynes like 14.

¹Department of Chemistry, University of Minnesota, 207 Pleasant Street, SE, Minneapolis, Minnesota 55455, USA.

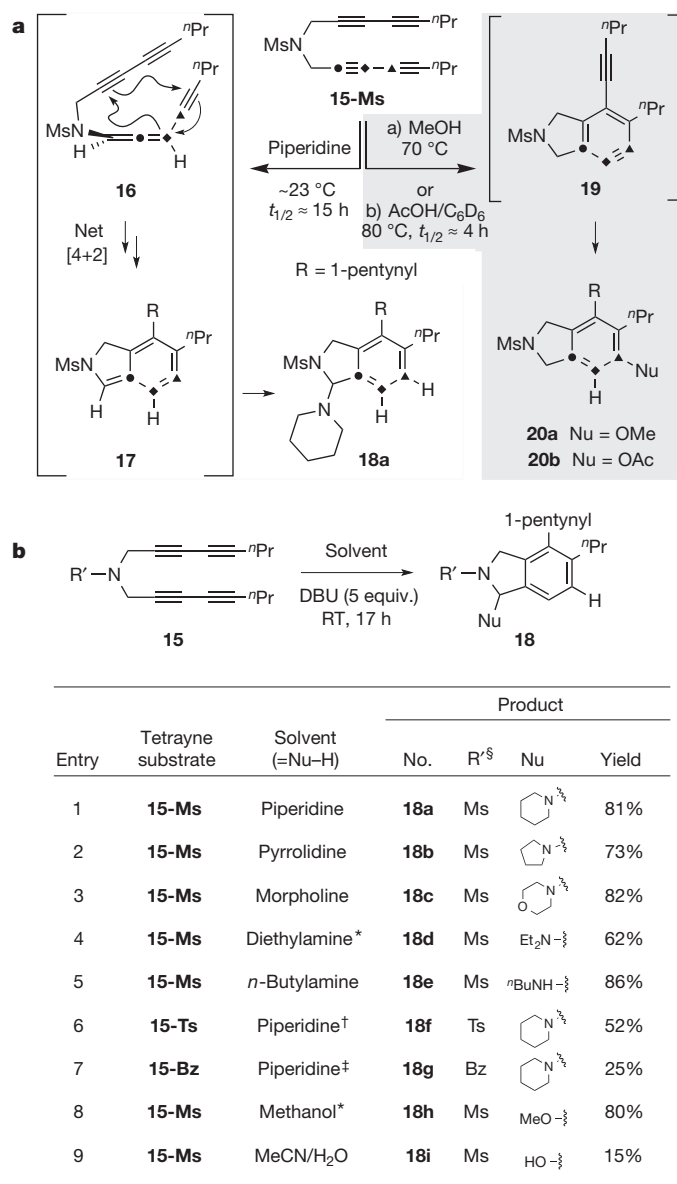


Figure 2 | PDDA cascades of tetraynes. **a**, The first example: substrate **15-Ms** undergoes base-promoted PDDA reaction and *in situ* trapping by piperidine to provide the adduct **18a**; the HDDA cyclization of **15-Ms** is slower. **b**, Examples indicating some of the scope of the PDDA cascades of substrates **15**. *Reaction performed at 40 °C for 17 h. †Reaction performed at 40 °C for 20 h. ‡Reaction performed using *t*BuOK (1 equiv.) in place of DBU as the added base at room temperature (RT) for 17 h. §Ms, CH₃SO₂; Ts, *p*-CH₃C₆H₄SO₂; Bz, PhCO.

intermediate, are generated by the cycloisomerization of allenynes like **14** (Fig. 1e) by the Myers–Saito reaction^{13–17} (Fig. 1e). However, we can find no previous examples of the generation of α ,3-dehydrotoluene derivatives by a formal cycloaddition event, the process we term here the pentadehydro-Diels–Alder (PDDA) reaction. We emphasize that in choosing this nomenclature, we do not mean to imply anything about the mechanism of this net [4 + 2] cycloaddition. Specifically, it should not be inferred that transformations like **4** + **11** to **12** are required to proceed by a concerted reaction pathway. Last, we note that α ,3-dehydrotoluene **12** is merely a tautomer of 3-methyl-1,2-dehydrobenzene (**9**). Accordingly, trapping of **12** leads to a toluene derivative that has been newly functionalized at its benzylic position, whereas capture of the tautomeric benzyne **9** gives rise to a newly substituted benzenoid product.

Our initial evidence for a PDDA process came from the reaction of tetrayne **15-Ms** in an ambient temperature solution of piperidine. The benzylic amine **18a** (Fig. 2a) was the only characterizable product formed in this experiment and was isolated in 81% yield following chromatographic purification. Tetrayne **15-Ms** was consumed with a half-life of approximately 15 h at room temperature. Based on several lines of evidence we have gathered and present below, we believe that the generation of **18a** is best described by (i) initial, rate-limiting piperidine-catalysed isomerization of **15-Ms** to produce the allenyne **16**, (ii) rapid PDDA cyclization to the dehydrotoluene **17**, and (iii) even more rapid trapping by protic piperidine of that reactive intermediate. We note that base-promoted isomerization of propargyl to allenyl sulfonamides is known¹⁸.

This reaction process is quite distinct from the course followed during an HDDA cascade (that is, sequential benzyne formation and trapping). Formation of a product in which one of the tethering atoms separating the diyne and diynophile has become functionalized has not been observed in any HDDA cascade⁷. We established that, in the absence of base, the tetrayne **15-Ms** is a well-behaved HDDA substrate, but only at elevated temperature (a half life of ~4 h at 80 °C in C₆D₆). We have trapped the resulting benzyne **19** with methanol^{19,20} or acetic acid^{5,19} to produce **20a** or **20b**, respectively (Fig. 2a). Taken together, these observations indicate that formation of the piperidine-trapped product **18a**, occurring at a substantially lower temperature, does not proceed via benzyne **19**. Thus, isomerization of the tetrayne **15-Ms** to the allene tautomer **16** occurs faster than does the thermal HDDA cyclization of **15-Ms**.

The non-nucleophilic base DBU (1,8-diazabicyclo[5.4.0]undec-7-ene) mildly accelerates the initial, rate-limiting isomerization of **15-Ms** to **16**. The rate of formation of adduct **18a** was approximately doubled (a *t*_{1/2} of 7 h versus 15 h) when five equivalents of DBU were added to the initial piperidine solution of **15-Ms**. Other secondary and primary amines participate in this transformation (Fig. 2b, entries 2–5). Other amides in the tether are also compatible (entries 6 and 7), although the reaction is slower with the benzamide **15-Bz**.

Oxygen nucleophiles will also trap the intermediate dehydrotoluene derivative **17** (Fig. 2b, entries 8 and 9). When carried out in methanol or aqueous acetonitrile, DBU-promoted PDDA reaction of the methanesulfonamide **15-Ms** gave the methoxylated or hydroxylated adducts **18h** or **18i**, respectively.

We have also achieved PDDA cyclizations with substrates in which the triply bonded acceptor moiety is a cyano rather than an alkynyl group (Fig. 3a). When dissolved in neat piperidine, diynyl nitriles **21** gave rise to the pyridine derivatives **22**, thereby establishing the viability of an aza-PDDA reaction. The ability of a cyano group to enter into the PDDA cycloisomerization is particularly noteworthy and decidedly distinct from its inertness to HDDA cyclization—neither we, nor others²¹, have observed nitriles engaging in that process. This is also the case for the nitriles **21**; in the absence of base, none gave evidence of cyclizing to a pyridine such as **23** (a 3,4-dehydropyridine), even upon heating to 150 °C in the presence of an excellent HDDA-aryne trap like acetic acid; only extensive decomposition was observed. Thus, the base-promoted tautomerization to **24** and subsequent PDDA cycloisomerization of that allene to the α ,3-dehydroazatoluene (or α ,3-dehydropicoline) intermediate **25** is considerably more facile than the HDDA cycloisomerization of its precursor tautomer **21**. This is entirely consistent with the observation presented earlier—namely, that the PDDA cyclization of **16** is much faster than the HDDA reaction of **15-Ms** (Fig. 2a).

As with the all-alkyne series already discussed (that is, **15**), the nitrile substrates **21** are also competent PDDA precursors when bearing a toluenesulfonyl, methanesulfonyl, or benzoyl electron-withdrawing group on the propargylic nitrogen atom (Fig. 3b). Again, amines (entries 1–4), water (entries 5, 6), and alcohols (entries 7–11) are all effective trapping nucleophiles.

We have performed several experiments (Fig. 4) that provide support for the proposed mechanism for these transformations. When either of

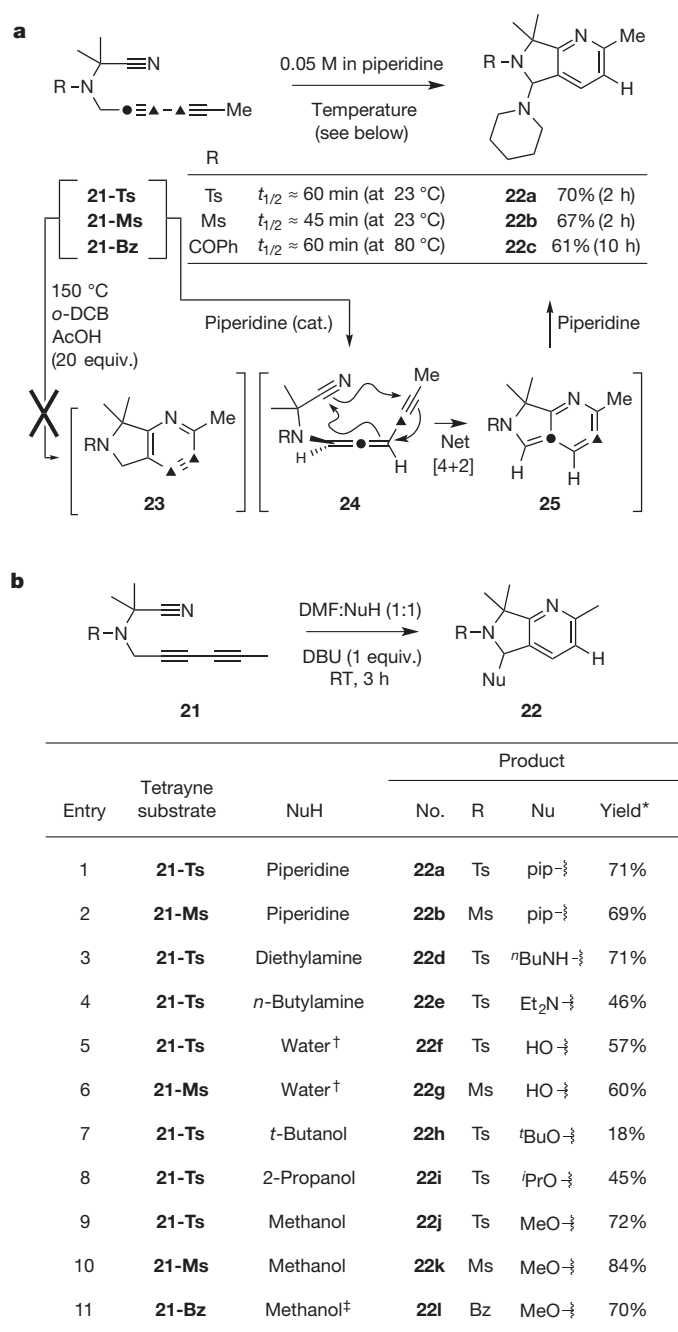


Figure 3 | Cyclizations of nitrile-containing diynes—the aza-PDDA.

a, Substrates **21**, non-competent reactants in HDDA cyclizations, undergo smooth, base-promoted PDDA reactions to give the piperidine-trapped adducts **22a–c**. **b**, Examples indicating some of the scope of the aza-PDDA cascade. *Yield of product following chromatographic purification.

[†]Reaction performed in a 3:1 (vol./vol.) mixture of CH_3CN :water. [‡]Reaction performed for 12 h.

the substrates **15-Ms** or **21-Ms** was incubated with DBU in a solution of deuteriochloroform (Fig. 4a), the major product (**26** or **27**, respectively) was a dichloroalkene in which the arene carbon corresponding to the *sp*-hybridized centre in the putative $\alpha,3$ -dehydrotoluene intermediate was fully deuterated. Presumably, initially formed trichloromethylated adducts like **28** underwent facile elimination induced by the excess DBU present.

In contrast to the behaviour of the geminal (gem)-dimethylated substrates **21** (see Fig. 3), the analogous substrate **29** (Fig. 4b) lacking those methyl groups gave a different outcome when held in a piperidine

solution at room temperature. No more than a trace amount of the expected PDDA cyclization product was observed. Instead, the enamine **31** was isolated as the principal product formed in this experiment. We presume that this arises by addition of the amine to the central carbon in allene **30** (to give a delocalized allylic/propargylic anion) rather than by direct hydroamination of the starting diyne **29**; we have performed a number of HDDA reactions on diyne substrates, not activated for tautomerization, in the presence of secondary amine trapping agents and never observed amination of a conjugated diyne. The different reaction course followed by **21** compared to that of **29** can be explained by the Thorpe–Ingold effect²²; the lack of the geminal substituents results in a widening of the bond angle and an increase in the distance (*r*) between the unsaturated centres in **30**. This presumably slows the rate of the PDDA cyclization, giving the piperidine time to intercept the allene. We have observed a similar phenomenon in the rate of HDDA cyclization of an analogous pair of triyne substrates²³ and also have found computational support for this interpretation (see below, Fig. 4c).

The energy diagram in Fig. 4c shows the relevant species involved in these PDDA reactions; these are the isomeric 1,3-diynes **32**, allenynes **33**, diradical intermediates **35** and $\alpha,3$ -dehydrotoluene/ $\alpha,3$ -dehydropicoline products **37**. Members of the **a/b** versus **c/d** series have an alkyne or a nitrile, respectively, as the pendant 2π -component. The diagram has been normalized so that each of the four PDDA reactant allenynes **33a–d** has the same free energy. The cyclizations proceed to the $\alpha,3$ -dehydro(aza)toluenes **37** by way of diradical intermediates **35**. Houk and co-workers have recently described a DFT analysis of (bimolecular) HDDA reactions and found that the diradical pathway is more favoured than the concerted pathway²⁴. We have observed the same for a series of intramolecular HDDA reactions, and found that the (U)B3LYP-D3BJ functional did an excellent job of correlating with our experimentally observed rates²⁵. We have not been successful in locating transition state structures corresponding to concerted PDDA reactions for **33a–d** at this level of theory.

Some notable points from these calculations are as follows: (i) the free energy differences between the 1,3-diynes **32a–d** and tautomeric allenynes **33a–d** are small, which serves as a reminder that the potential energies of the participating functional groups in **33a–d** are also high, comparable to those in **32a–d**; (ii) as with triyne-to-benzynes conversions, the overall energies of reaction from **33a,b** to the reactive $\alpha,3$ -dehydrotoluenes **37a,b** are exergonic (by an amount ΔG° of $\sim 35 \text{ kcal mol}^{-1}$), although not to as large an extent as those computed for HDDA cyclizations to benzyne (about $-50 \text{ kcal mol}^{-1}$)^{5,26}; (iii) the corresponding energy differences between the nitrile-containing allenynes **33c,d** and the product $\alpha,3$ -dehydroazatoluenes **37c,d** are even smaller, reflecting the inherently lower potential energy of a nitrile triple bond versus that of an alkyne; (iv) the magnitude of the computed free energy of activation (ΔG^\ddagger) values for the first (and slower) step in the PDDA cyclization (see **34** (TS1) in Fig. 4c) are not inconsistent with the fact that our PDDA cyclizations are proceeding rapidly at less than near-ambient temperatures; and (v) the difference in the computed ΔG^\ddagger for the first step in the PDDA reaction of the nitrile **33c** versus that of **33d** ($20.6 \text{ kcal mol}^{-1}$ versus $16.8 \text{ kcal mol}^{-1}$) is consistent with the differing behaviour of allenyne **24** (Fig. 3a) versus the aza-analogue **30** (Fig. 4b). We recall that the former underwent smooth PDDA cyclization to **25** en route to the piperidine-trapped adduct **22b** (Fig. 3a), whereas the latter cyclized so slowly that interception by piperidine occurred to produce the enamine **31**. The optimized geometries computed for the reactive conformers of **33c** versus **33d** (without versus with gem-dimethyl groups²²) differ substantially (0.21 \AA) in the distance (*r*) between the nitrile and central allene carbons, indicating that the latter is better poised for surmounting the TS1 activation barrier.

Last, we devised an experiment to unambiguously demonstrate the intermediacy of an allenyne (see **40**, Fig. 4d). The hydroxytetrayne **38** reacted readily with chlorodiphenylphosphine to produce

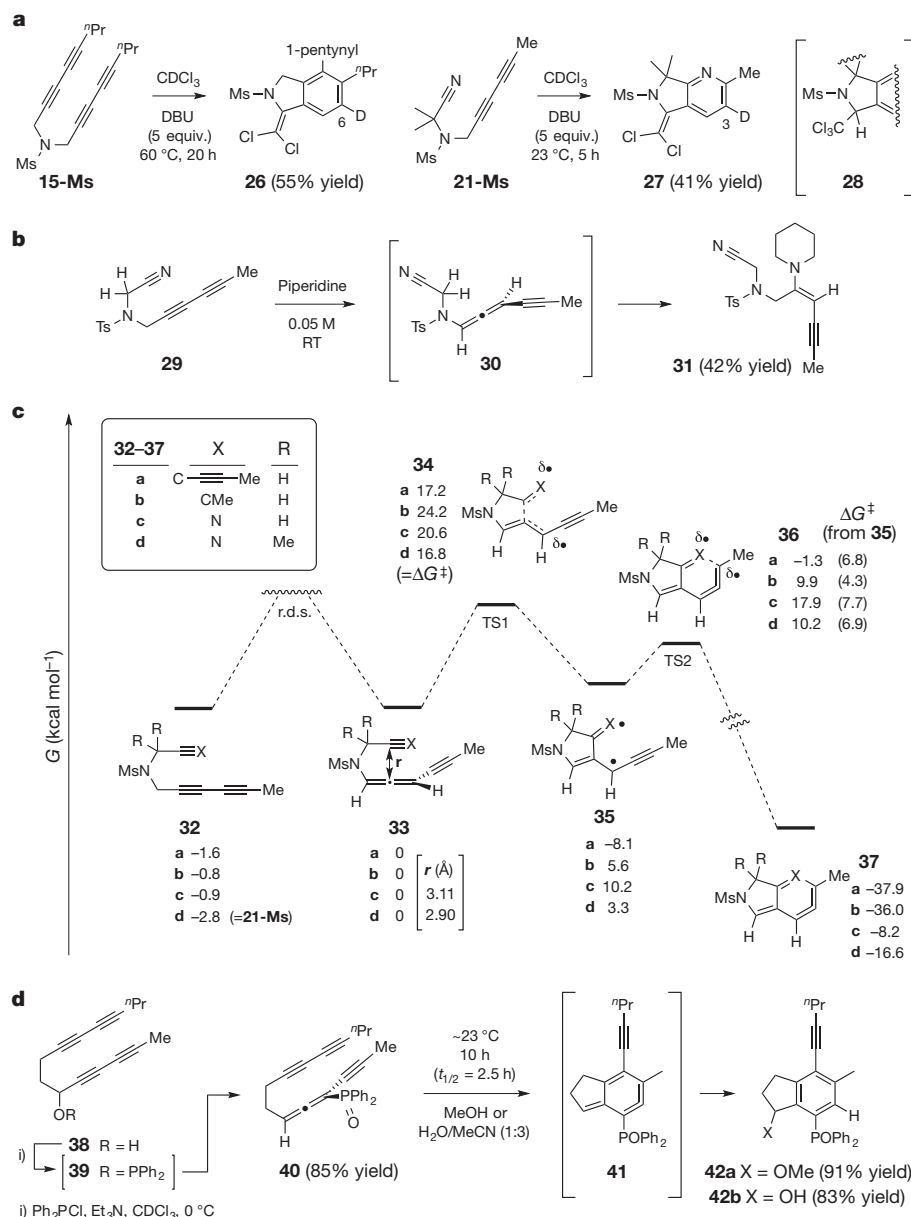


Figure 4 | Mechanistic aspects of the PDPA reaction. **a, b**, Indirect evidence for intermediacy of $\alpha,3$ -dehydrotoluenes and allenes. **c**, Relative free energies (G) from DFT calculations ((U)B3LYP-D3BJ/6-311+G(d,p); a solvation model (using Et_2NH) was employed (see Supplementary Information)) of the directly relevant minima (32, 33, 35, and 37) and two transition state structures (34 (TS1) and 36 (TS2)) for the PDPA

cyclization via the diradical intermediate 35. Values beside 'a–d' for each of 32–37 are the computed energies (G) in kcal mol^{-1} . The $2\pi \text{ C}\equiv\text{X}$ component is either an alkyne (**a, b**) or a nitrile (**c, d**). r.d.s., rate-determining step. **d**, An (isolable) allenyne, the phosphine oxide 40, readily cyclizes to the benzenoid product 42.

the transient phosphinite 39, which smoothly rearranged^{14,27} at sub-ambient temperature to the allenyldiphenylphosphine oxide 40. This labile compound was observed to degrade upon handling at room temperature, but could be rapidly purified and characterized. Dissolving 40 in methanol or water/acetonitrile gave rise spontaneously to the corresponding trapped product 42a or 42b, respectively. We view this as strong support for the intermediacy of 41 and the PDPA mechanism posited here throughout. That these reactions proceed smoothly in the absence of base also argues against a mechanism initiated by nucleophilic attack on the allene terminal carbon of the transient intermediates 16 (Fig. 2a) or 24 (Fig. 3a). The known reaction of methanol with Myers–Saito reaction-derived $\alpha,3$ -dehydrotoluenes in similar fashion is also relevant^{28–30}.

In summary, we have described a new class of reaction—the formal [4 + 2] cycloaddition of an allenyne with a pendant alkyne

(or nitrile) to produce an $\alpha,3$ -dehydro(aza)toluene derivative. We call this a pentadehydro-Diels–Alder (or PDPA) reaction. In the majority of examples reported, the base-promoted isomerization of a precursor N -1,3-diynyl sulfonamide to the reactive allenyl tautomer is overall rate-limiting. The PDPA then proceeds rapidly, much more quickly than the HDDA cyclization of the precursor 1,3-diyne. We also have shown that the PDPA-derived dehydrotoluene, itself a reactive intermediate, can be trapped by a variety of N -, O -, and C -centred nucleophiles (Figs 2b and 3b). In one instance the conjugated allenyne has been isolated (40, Fig. 4d) and its facile cyclization and *in situ* trapping observed. Finally, nitriles, which do not participate in HDDA reactions, now enter the realm of reactivity, resulting in pyridine-containing products (Fig. 3).

Received 7 December 2015; accepted 12 February 2016.

Published online 18 April 2016.

- Diels, O. & Alder, K. Synthesen in der hydroaromatischen Reihe. *Justus Liebigs Ann. Chem.* **460**, 98–122 (1928).
- Wessig, P. & Müller, G. The dehydro-Diels-Alder reaction. *Chem. Rev.* **108**, 2051–2063 (2008).
- Bradley, A. Z. & Johnson, R. P. Thermolysis of 1,3,8-nonatriyne: evidence for intramolecular [2+4] cycloaromatization to a benzyne intermediate. *J. Am. Chem. Soc.* **119**, 9917–9918 (1997).
- Miyawaki, K., Suzuki, R., Kawano, T. & Ueda, I. Cycloaromatization of a non-conjugated polyenyne system: synthesis of 5H-benzo[d]fluoreno[3,2-b]pyrans via diradicals generated from 1-[2-{4-(2-alkoxymethylphenyl)butan-1,3-diynyl}]phenylpentan-2,4-diyn-1-ols and trapping evidence for the 1,2-didehydrobenzene diradical. *Tetrahedr. Lett.* **38**, 3943–3946 (1997).
- Hoye, T. R., Baire, B., Niu, D., Willoughby, P. H. & Woods, B. P. The hexadehydro-Diels-Alder reaction. *Nature* **490**, 208–212 (2012).
- Yun, S. Y., Wang, K.-P., Lee, N.-K., Mamidipalli, P. & Lee, D. Alkane C–H insertion by aryne intermediates with a silver catalyst. *J. Am. Chem. Soc.* **135**, 4668–4671 (2013).
- Holden (née Hall), C. & Greaney, M. F. The hexadehydro-Diels-Alder reaction: a new chapter in aryne chemistry. *Angew. Chem. Int. Ed.* **53**, 5746–5749 (2014).
- Wang, K.-P., Yun, S. Y., Mamidipalli, P. & Lee, D. Silver-mediated fluorination, trifluoromethylation, and trifluoromethylthiolation of arynes. *Chem. Sci.* **4**, 3205–3211 (2013).
- Niu, D. & Hoye, T. R. The aromatic ene reaction. *Nature Chem.* **6**, 34–40 (2014).
- Bergman, R. G. Reactive 1,4-dehydroaromatics. *Acc. Chem. Res.* **6**, 25–31 (1973).
- Johnson, R. P. Dehydropericyclic routes to reactive intermediates. *J. Phys. Org. Chem.* **23**, 283–292 (2010).
- Logan, C. F., Ma, J. C. & Chen, P. The photoelectron spectrum of the α ,3-dehydrotoluene biradical. *J. Am. Chem. Soc.* **116**, 2137–2138 (1994).
- Myers, A. G., Kuo, E. Y. & Finney, N. S. Thermal generation of α ,3-dehydrotoluene from (Z)-1,2,4-heptatrien-6-yne. *J. Am. Chem. Soc.* **111**, 8057–8059 (1989).
- Nagata, R., Yamanaka, H., Okazaki, E. & Saito, I. Biradical formation from acyclic conjugated eneyne-allene system related to neocarzinostatin and esperamicin-calicheamicin. *Tetrahedr. Lett.* **30**, 4995–4998 (1989).
- Mohamed, R. K., Peterson, P. W. & Alabugin, I. V. Concerted reactions that produce diradicals and zwitterions: electronic, steric, conformational, and kinetic control of cycloaromatization processes. *Chem. Rev.* **113**, 7089–7129 (2013).
- Sakai, T. & Danheiser, R. L. Cyano Diels-Alder and cyano ene reactions. applications in a formal [2+2+2] cycloaddition strategy for the synthesis of pyridines. *J. Am. Chem. Soc.* **132**, 13203–13205 (2010).
- Karmakar, R., Yun, S. Y., Chen, J., Xia, Y. & Lee, D. Benzannulation of triynes to generate functionalized arenes by spontaneous incorporation of nucleophiles. *Angew. Chem. Int. Ed.* **54**, 6582–6586 (2015).
- Grigg, R. & Sansano, J. M. Sequential hydrostannylation-cyclisation of δ - and ω -allenyl aryl halides. Cyclisation at the proximal carbon. *Tetrahedron* **52**, 13441–13454 (1996).
- Karmakar, R., Yun, S. Y., Wang, K.-P. & Lee, D. Regioselectivity in the nucleophile trapping of arynes: the electronic and steric effects of nucleophiles and substituents. *Org. Lett.* **16**, 6–9 (2014).
- Willoughby, P. H. *et al.* Mechanism of the reactions of alcohols with o-benzynes. *J. Am. Chem. Soc.* **136**, 13657–13665 (2014).
- Kimura, H., Torikai, K., Miyawaki, K. & Ueda, I. Scope of the thermal cyclization of nonconjugated ene-yne-nitrile system: a facile synthesis of cyanofluorene derivatives. *Chem. Lett.* **37**, 662–663 (2008).
- Beesley, R. M., Ingold, C. K. & Thorpe, J. F. The formation and stability of spiro-compounds. Part I: Spiro-compounds from cyclohexane. *J. Am. Chem. Soc.* **107**, 1081–1092 (1915).
- Woods, B. P., Baire, B. & Hoye, T. R. Rates of hexadehydro-Diels-Alder (HDDA) cyclizations: impact of the linker structure. *Org. Lett.* **16**, 4578–4581 (2014).
- Liang, Y., Hong, X., Yu, P. & Houk, K. N. Why alkynyl substituents dramatically accelerate hexadehydro-Diels-Alder (HDDA) reactions: stepwise mechanisms of HDDA cycloadditions. *Org. Lett.* **16**, 5702–5705 (2014).
- Marell, D. J. *et al.* Mechanism of the intramolecular hexadehydro-Diels-Alder reaction. *J. Org. Chem.* **80**, 11744–11754 (2015).
- Ajaz, A. *et al.* Concerted vs. stepwise mechanisms in dehydro-Diels-Alder reactions. *J. Org. Chem.* **76**, 9320–9328 (2011).
- Schmitt, M., Strittmatter, M., Vollmann, K. & Kiau, S. Intramolecular formal Diels-Alder reaction in enyne allenes. A new synthetic route to benzofluorenes and indeno[1,2-g]quinolines. *Tetrahedr. Lett.* **37**, 999–1002 (1996).
- Myers, A. G., Dragovich, P. S. & Kuo, E. Y. Studies on the thermal generation and reactivity of a class of (σ , π)-1,4-biradicals. *J. Am. Chem. Soc.* **114**, 9369–9386 (1992).
- Cremins, M. E., Hughes, T. S. & Carpenter, B. K. Mechanistic studies on the cyclization of (Z)-1,2,4-heptatrien-6-yne in methanol: a possible nonadiabatic thermal reaction. *J. Am. Chem. Soc.* **127**, 6652–6661 (2005).
- Peterson, P. W., Mohamed, R. K. & Alabugin, I. V. How to lose a bond in two ways — the diradical/zwitterion dichotomy in cycloaromatization reactions. *Eur. J. Org. Chem.* 2505–2527 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements Financial support from the National Institute of General Medical Sciences (GM65597) and the National Cancer Institute (CA76497) of the US Department of Health and Human Services is acknowledged. Portions of this work were performed using resources available through the University of Minnesota Supercomputing Institute (MSI). NMR spectra were recorded using instrumentation purchased with funds from the NIH Shared Instrumentation Grant programme (S10OD011952). We thank D. J. Marell for guidance in several aspects of the computations.

Author Contributions T.W. and R.R.N. carried out the experiments and contributed equally to the overall work; S.K.T. performed the computational studies. All authors interpreted the results and prepared the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.R.H. (hoye@umn.edu).

Rapid cycling of reactive nitrogen in the marine boundary layer

Chunxiang Ye¹, Xianliang Zhou^{1,2}, Dennis Pu², Jochen Stutz³, James Festa³, Max Spolaor³, Catalina Tsai³, Christopher Cantrell⁴, Roy L. Mauldin III^{4,5}, Teresa Campos⁶, Andrew Weinheimer⁶, Rebecca S. Hornbrook⁶, Eric C. Apel⁶, Alex Guenther⁷, Lisa Kaser⁶, Bin Yuan^{8,9}, Thomas Karl¹⁰, Julie Haggerty⁶, Samuel Hall⁶, Kirk Ullmann⁶, James N. Smith^{6,11}, John Ortega⁶ & Christoph Knöche^{6†}

Nitrogen oxides are essential for the formation of secondary atmospheric aerosols and of atmospheric oxidants such as ozone and the hydroxyl radical, which controls the self-cleansing capacity of the atmosphere¹. Nitric acid, a major oxidation product of nitrogen oxides, has traditionally been considered to be a permanent sink of nitrogen oxides¹. However, model studies predict higher ratios of nitric acid to nitrogen oxides in the troposphere than are observed^{2,3}. A 'renoxification' process that recycles nitric acid into nitrogen oxides has been proposed to reconcile observations with model studies^{2–4}, but the mechanisms responsible for this process remain uncertain^{5–9}. Here we present data from an aircraft measurement campaign over the North Atlantic Ocean and find evidence for rapid recycling of nitric acid to nitrous acid and nitrogen oxides in the clean marine boundary layer via particulate nitrate photolysis. Laboratory experiments further demonstrate the photolysis of particulate nitrate collected on filters at a rate more than two orders of magnitude greater than that of gaseous nitric acid, with nitrous acid as the main product. Box model calculations based on the Master Chemical Mechanism^{10,11} suggest that particulate nitrate photolysis mainly sustains the observed levels of nitrous acid and nitrogen oxides at midday under typical marine boundary layer conditions. Given that oceans account for more than 70 per cent of Earth's surface, we propose that particulate nitrate photolysis could be a substantial tropospheric nitrogen oxide source. Recycling of nitrogen oxides in remote oceanic regions with minimal direct nitrogen oxide emissions could increase the formation of tropospheric oxidants and secondary atmospheric aerosols on a global scale.

Nitrogen oxides ($\text{NO}_x = \text{NO}$ and NO_2) are essential in the formation of secondary aerosol and atmospheric oxidants, such as ozone (O_3) and hydroxyl radicals (OH)¹. Nitric acid (HNO_3) is a major oxidation product of NO_x . The formation of HNO_3 has traditionally been considered to be a permanent sink of NO_x owing to the small photolysis frequency of gaseous HNO_3 and to its rapid removal from the troposphere by wet and dry deposition¹. However, this traditional view has recently been challenged by laboratory studies demonstrating that photolysis of surface-adsorbed HNO_3 is enhanced by one to four orders of magnitude, compared to gaseous HNO_3 and aqueous nitrate, with NO_x and nitrous acid (HONO) as products^{5–8}. Field observations have provided further evidence that photolysis of HNO_3 adsorbed on the forest canopy surfaces is the primary daytime HONO source for the overlying atmosphere⁹. Model simulations that include HONO observations or photolytic HONO formation mechanisms have substantially improved estimates of atmospheric oxidant formation^{12,13}.

However, it remains unknown whether photolysis of particulate nitrate (pNO_3) contributes to the HONO and NO_x budgets and affects the formation of oxidants and secondary aerosol.

The Nitrogen, Oxidants, Mercury and Aerosol Distributions, Sources and Sinks (NOMADSS) field study was conducted in the summer of 2013 aboard the NSF/NCAR C-130 aircraft. One objective was to investigate the role of pNO_3 photolysis in reactive nitrogen cycling. Two research flights in particular, RF14 on 5 July 2013 and RF16 on 8 July 2013, involved similar flight tracks in the marine boundary layer (MBL) and the free troposphere over the North Atlantic Ocean (flight tracks from $33^\circ 39' \text{N}$, $77^\circ 42' \text{W}$ to $32^\circ 11' \text{N}$, $77^\circ 41' \text{W}$). The flight periods in the MBL and the free troposphere over the Atlantic Ocean took place between 12:00 and 15:00 Eastern Standard Time (equivalent to Coordinated Universal Time (UTC) minus 5 h). The solar elevation angle during the flight legs in the MBL for both research flights was between 17° and 31° . A comprehensive suite of parameters relevant to the research objective were simultaneously measured, including HONO , NO_x , pNO_3 , HNO_3 , aerosol surface area density (for particle diameters $< 1 \mu\text{m}$), OH , hydroperoxyl (HO_2) and alkylperoxyl (RO_2) radicals, bromine oxide (BrO), iodine oxide (IO), O_3 , volatile organic compounds (VOCs), and photolysis frequencies J (see Extended Data Table 1). The air masses encountered in the MBL consisted mostly of aged marine air circulating over the North Atlantic Ocean under a Bermuda high-pressure system as indicated by back-trajectories, with slight influence from coastal emissions in RF16 (Extended Data Fig. 1).

The observed mixing ratios of HONO , NO_x and pNO_3 versus altitude are shown in Fig. 1. The mixing ratios of NO_x ranged from 10 to 40 parts per trillion by volume (p.p.t.v.), and were comparable to previous observations in the clean MBL¹⁴. The mean mixing ratios of HONO (± 1 s.d., standard deviation) within the MBL were 11.3 ± 1.6 p.p.t.v. for RF14 and 8.8 ± 2.3 p.p.t.v. for RF16. These HONO values are substantially lower than other airborne observations in the continental boundary layer over Northern Michigan ($8\text{--}70$ p.p.t.v.)¹⁵ and in the morning residual layer over an industrial region of Northern Italy (up to 150 p.p.t.v.)¹⁶. These low HONO values are consistent with the lower levels of its precursors, such as NO_x , in this clean MBL compared to the other two locations. In addition, the ocean surface is likely to be a net HONO sink given the slight alkalinity of sea water, while the land surface is typically a net HONO source for the overlying atmosphere^{9,15}.

The photolysis lifetime of HONO during the flight legs discussed here was approximately 12 min. To balance the photolytic loss and maintain a nearly steady-state HONO concentration, there must be an *in situ* HONO source well distributed in the MBL. NO_x is a well

¹Wadsworth Center, New York State Department of Health, Albany, New York, USA. ²Department of Environmental Health Sciences, State University of New York, Albany, New York, USA.

³Department of Atmospheric and Oceanic Sciences, University of California, Los Angeles (UCLA), California, USA. ⁴Department of Atmospheric and Oceanic Sciences, University of Colorado at Boulder, Boulder, Colorado, USA. ⁵Department of Physics, University of Helsinki, Helsinki, Finland. ⁶National Center for Atmospheric Research, Boulder, Colorado, USA. ⁷Pacific Northwest National Laboratory, Richland, Washington, USA. ⁸NOAA, Earth System Research Laboratory, Chemical Sciences Division, Boulder, Colorado, USA. ⁹Cooperative Institute for Research in Environmental Sciences, University of Colorado at Boulder, Boulder, Colorado, USA. ¹⁰Institute for Meteorology and Geophysics, University of Innsbruck, Innsbruck, Austria. ¹¹University of Eastern Finland, Kuopio, Finland. [†]Present address: Meteorologisches Institut, Ludwig-Maximilians-Universität München, Germany.

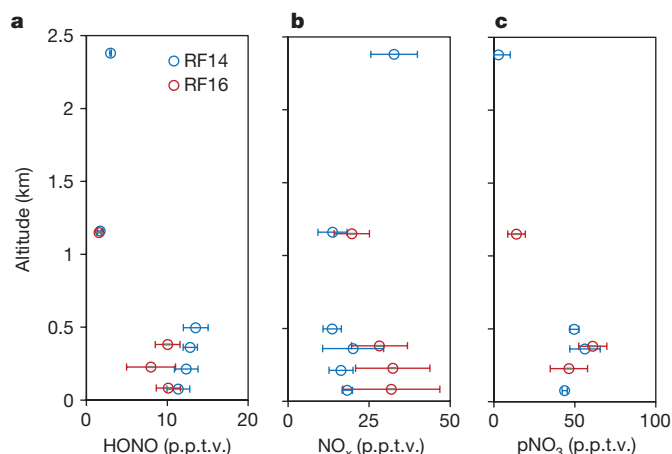


Figure 1 | Vertical profiles of HONO, NO_x and pNO_3 in marine air off the coast of North and South Carolina on 5 July 2013 (RF 14) and 8 July 2013 (RF 16). a, HONO; b, NO_x ; c, pNO_3 . Each error bar represents ± 1 s.d. of the measurements. The height of the MBL is about 1 km.

known HONO precursor, especially in urban environments¹⁷. However, there was no significant and positive correlation ($r^2 < 0.1$) between the concentrations of HONO and NO_x , suggesting that NO_x was not an important HONO precursor in the marine atmosphere. On the other hand, HONO and pNO_3 concentrations were much higher in the MBL than in the free troposphere (Fig. 1) and were significantly correlated ($r^2 = 0.62$ for RF14 and $r^2 = 0.61$ for RF16), suggesting that pNO_3 was a potential HONO precursor in the marine atmosphere. Indeed, the HONO source required to sustain the observed HONO concentration correlated well with the product of the pNO_3 concentration and HNO_3 photolysis frequency ($r^2 = 0.82$ for RF14 and $r^2 = 0.80$ for RF16) (Fig. 2). An enhancement factor of about 300 can be calculated from the slopes of linear least-squares fits for the photolysis rate constant of pNO_3 relative to that of gaseous HNO_3 , if pNO_3 photolysis is the *in situ* HONO source sustaining the observed HONO concentrations after removing the minor contributions from NO_x -related reactions. A normalized photolysis rate constant for pNO_3 , $J_{\text{pNO}_3}^N$, of $\sim 2.0 \times 10^{-4} \text{ s}^{-1}$ can be derived from the enhancement factor, scaled to the typical tropical summer conditions on the ground (solar elevation angle $\theta = 0^\circ$), corresponding to a gaseous HNO_3 photolysis rate constant of $\sim 7.0 \times 10^{-7} \text{ s}^{-1}$ (refs 1 and 18).

To confirm the role of pNO_3 photolysis as a HONO source, aerosol samples were collected on board the C-130 aircraft using Teflon filters for the laboratory determination of the photolysis rate constants of pNO_3 . A $J_{\text{pNO}_3}^N$ value of $1.0 \times 10^{-4} \text{ s}^{-1}$ was determined from one aerosol sample collected during the MBL flight RF16. In addition, photolysis rate constants were determined to be in the range $1.3 \times 10^{-5} \text{ s}^{-1}$ to $3.1 \times 10^{-4} \text{ s}^{-1}$, with a median of $1.3 \times 10^{-4} \text{ s}^{-1}$ and a mean ± 1 s.d. of $(1.7 \pm 1.1) \times 10^{-4} \text{ s}^{-1}$, from seven filter samples collected in various air masses during the NOMADSS flights, including one over the ocean (RF 16) and six over terrestrial environments in the southern and southeastern USA. HONO was the major product of pNO_3 photolysis under the experimental conditions, with an average HONO/ NO_x production ratio of ~ 2 . The laboratory-determined values are in reasonably good agreement with the value of $\sim 2.0 \times 10^{-4} \text{ s}^{-1}$ inferred from the field observations (Fig. 2).

Both the field-inferred and the laboratory-determined $J_{\text{pNO}_3}^N$ values are also well within a reported range for the surface HNO_3 photolysis rate constant, from $2.2 \times 10^{-5} \text{ s}^{-1}$ on glass surfaces^{5,19} to $1.2 \times 10^{-3} \text{ s}^{-1}$ on urban grime⁶. Similar to surface-adsorbed HNO_3 , the high photolysis rate of pNO_3 may be due in part to its enhanced absorption cross-section, compared to that of gaseous HNO_3 (refs 7 and 8). The photolysis rate of pNO_3 associated with sea-salt aerosol may also be enhanced in the unique chemical environment of the MBL. Organic matter is ubiquitous and highly enriched in sea-salt

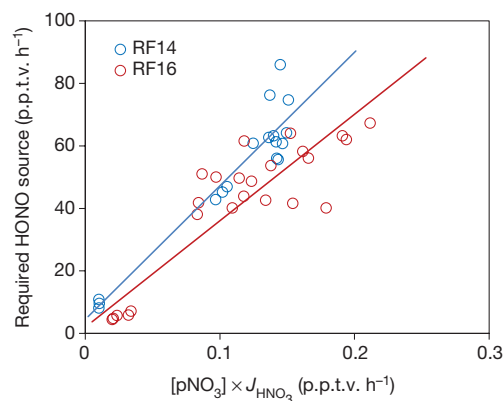


Figure 2 | Correlation between the required *in situ* HONO source and the product of the pNO_3 concentration and the photolysis frequency of gaseous HNO_3 . The lines are the linear least-squares fits to the data. The correlation coefficients (r^2) are 0.82 and 0.80 for RF14 and RF16, respectively. The relative uncertainty (± 1 s.d.) is within 25% for the required *in situ* HONO source and within 35% for $J_{\text{HNO}_3} \times [\text{pNO}_3]$ in the MBL, and is within $\sim 50\%$ and $\sim 70\%$ in the free troposphere.

aerosols²⁰ and may serve as a photosensitizer for the photolysis of pNO_3 . Furthermore, the relative humidity was measured to be mostly higher than the deliquescence point of sea-salt particles, and thus sea-salt particles were in the form of liquid droplets in the MBL. The light intensity is enhanced in sea-salt droplets owing to resonance and refraction²¹, and the nitrate ion is attracted by Br^- to the air–water interface, where it becomes more photochemically reactive²².

All the evidence from the field and the laboratory indicate that fast photolysis of pNO_3 in sea-salt aerosol sustains a majority of the observed HONO level in the MBL. Since HONO is readily photolysed to NO and OH during the day, the photolysis of pNO_3 becomes a ‘renoxification’ pathway. Sea-salt particles scavenge HNO_3 in the MBL, converting photochemically inert HNO_3 into photochemically reactive pNO_3 and thus accelerating the photochemical recycling of nitric acid/nitrate to NO_x in the MBL. Figure 3 shows diurnal cycling of the pNO_3 –HONO– NO_x system, simulated by a box model based on the Master Chemical Mechanism (MCM v3.2)^{10,11}. The observed and the predicted values during the early afternoon are in good agreement, within the measurement and model uncertainties. The simulated HONO and NO_x diurnal profiles were similar to those observed in the polar regions, where the photolysis of snowpack nitrate was the dominant HONO and NO_x source²³ and to that of NO_x in the clean MBL, where a photolytic NO_x source was proposed^{24,25}.

The predicted daytime HONO maximum is due to the strong HONO source from pNO_3 photolysis (Extended Data Fig. 2a). Other HONO sources included in our box model are the gas-phase reactions of OH with NO, of excited NO_2 with H_2O (refs 26 and 27), and of HO_2 – H_2O with NO_2 (refs 16 and 28), and heterogeneous reactions of NO_2 on sea-salt aerosol particles, but all of these have been found to be insignificant. The predicted daytime NO_x maximum is also due to the strong NO_x source from pNO_3 photolysis, partly though HONO as an intermediate (Extended Data Fig. 2b). Other known ‘renoxification’ processes in the gas phase, such as the $\text{HNO}_3 + \text{OH}$ reaction and HNO_3 photolysis, are negligible NO_x sources in the clean MBL. NO_x is effectively oxidized to HNO_3 and pNO_3 via several mechanisms at comparable rates (Extended Data Fig. 2b). Of these mechanisms, the gas-phase reactions of NO_2 with OH and XO (X is primarily Br and I) are important HNO_3 and pNO_3 sources in the MBL²⁹; the gas-phase reactions of NO with RO_2 radicals leading to organic nitrates have recently been reported to be the main sink of NO_x in low- NO_x forested environments³⁰. The total nitrate (that is, $\text{HNO}_3 + \text{pNO}_3$) shows a net loss in the morning and a net gain from the mid-afternoon to early evening, with a balanced daily budget (Fig. 3). Model simulation results indicate that the pNO_3 – NO_x cycling occurs on a scale of a few

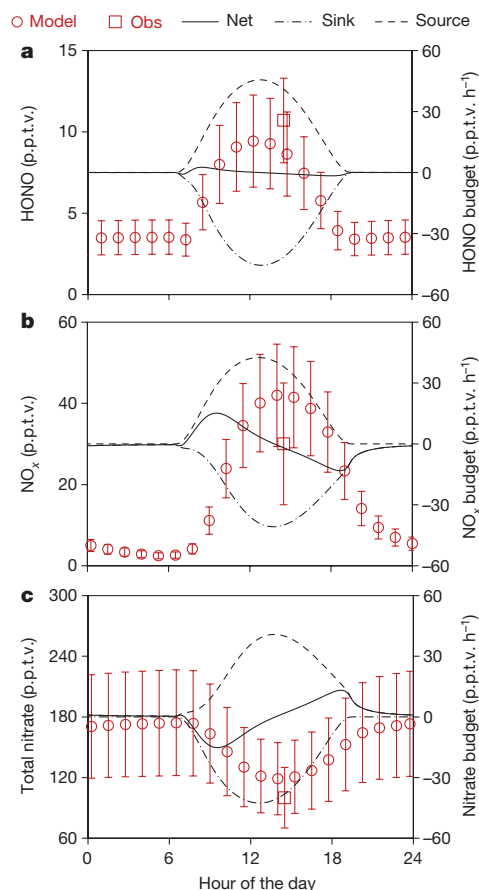


Figure 3 | Simulated diurnal profiles of mixing ratios and budgets for HONO, NO_x and total nitrate in the MBL. a, HONO; b, NO_x ; c, total nitrate ($\text{HNO}_3 + \text{pNO}_3$). HONO sources include reactions of excited NO_2 with H_2O (refs 26 and 27), of $\text{H}_2\text{O}-\text{HO}_2$ with NO_2 (refs 16 and 28), of OH with NO, the heterogeneous reaction of NO_2 on aerosols, and the photolysis of pNO_3 . NO_x sinks consist mainly of reactions of RO_2 with NO (ref. 30), of XO (primarily BrO and IO) with NO_2 (ref. 29) and of OH with NO_2 . The error bar represents ± 1 s.d. of the model calculations and field observations.

hours during the day, and is capable of sustaining the observed midday levels of HONO and NO_x in the MBL.

The rapid pNO_3 –HONO– NO_x cycling suggests that the total nitrate in the troposphere is a NO_x reservoir instead of the permanent NO_x sink. The traditional view of the reactive nitrogen budget needs revision. Oceans account for over 70% of Earth's surface, so this recycling NO_x could contribute substantially to the global NO_x source. Moreover, this recycling source of NO_x occurs in the MBL, where direct NO_x emissions are negligible, and thus it could have profound impacts on global tropospheric chemistry, such as oxidant and secondary aerosol formation.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 14 July 2015; accepted 28 January 2016.

Published online 11 April 2016.

1. Finlayson-Pitts, B. J. & Pitts, J. N., Jr. *Chemistry of the Upper and Lower Atmosphere: Theory, Experiments and Applications* (Academic, 2000).
2. Wang, K., Zhang, Y., Nenes, A. & Fountoukis, C. Implementation of dust emission and chemistry into the Community Multiscale Air Quality modeling system and initial application to an Asian dust storm episode. *Atmos. Chem. Phys.* **12**, 10209–10237 (2012).
3. Deng, J., Wang, T., Liu, L. & Jiang, F. Modeling heterogeneous chemical processes on aerosol surface. *Particology* **8**, 308–318 (2010).
4. Kumar, R. *et al.* Effects of dust aerosols on tropospheric chemistry during a typical pre-monsoon season dust storm in northern India. *Atmos. Chem. Phys.* **14**, 6813–6834 (2014).

5. Zhou, X. *et al.* Nitric acid photolysis on surfaces in low- NO_x environments: significant atmospheric implications. *Geophys. Res. Lett.* **30**, 2217 (2003).
6. Baergen, A. M. & Donaldson, D. J. Photochemical renoxification of nitric acid on real urban grime. *Environ. Sci. Technol.* **47**, 815–820 (2013).
7. Zhu, C., Xiang, B., Chu, L. T. & Zhu, L. 308 nm photolysis of nitric acid in the gas phase, on aluminum surfaces, and on ice films. *J. Phys. Chem. A* **114**, 2561–2568 (2010).
8. Du, J. & Zhu, L. Quantification of the absorption cross sections of surface-adsorbed nitric acid in the 335–365 nm region by Brewster angle cavity ring-down spectroscopy. *Chem. Phys. Lett.* **511**, 213–218 (2011).
9. Zhou, X. *et al.* Nitric acid photolysis on forest canopy surface as a source for tropospheric nitrous acid. *Nature Geosci.* **4**, 440–443 (2011).
10. Saunders, S. M., Jenkin, M. E., Derwent, R. G. & Pilling, M. J. Protocol for the development of the Master Chemical Mechanism, MCM v3 (part A): tropospheric degradation of non-aromatic volatile organic compounds. *Atmos. Chem. Phys.* **3**, 161–180 (2003).
11. Jenkin, M. E., Saunders, S. M., Wagner, V. & Pilling, M. J. Protocol for the development of the Master Chemical Mechanism, MCM v3 (part B): tropospheric degradation of aromatic volatile organic compounds. *Atmos. Chem. Phys.* **3**, 181–193 (2003).
12. Ren, X. *et al.* OH, HO_2 , and OH reactivity during the PMTACS–NY Whiteface Mountain 2002 campaign: observations and model comparison. *J. Geophys. Res.* **111**, D10S03 (2006).
13. Czader, B. H. *et al.* Modeling nitrous acid and its impact on ozone and hydroxyl radical during the Texas Air Quality Study 2006. *Atmos. Chem. Phys.* **12**, 6939–6951 (2012).
14. Lee, J. D. *et al.* Year-round measurements of nitrogen oxides and ozone in the tropical North Atlantic marine boundary layer. *J. Geophys. Res.* **114**, D21302 (2009).
15. Zhang, N. *et al.* Aircraft measurement of HONO vertical profiles over a forested region. *Geophys. Res. Lett.* **36**, L15820 (2009).
16. Li, X. *et al.* Missing gas-phase source of HONO inferred from Zeppelin measurements in the troposphere. *Science* **344**, 292–296 (2014).
17. Kleffmann, J. Daytime sources of nitrous acid (HONO) in the atmospheric boundary layer. *ChemPhysChem* **8**, 1137–1144 (2007).
18. Jankowski, J. J., Kieber, D. J., Mopper, K. & Neale, P. J. Development and intercalibration of ultraviolet solar actinometers. *Photochem. Photobiol.* **71**, 431–440 (2000).
19. Ramazan, K. A., Syomin, D. & Finlayson-Pitts, B. J. The photochemical production of HONO during the heterogeneous hydrolysis of NO_2 . *Phys. Chem. Chem. Phys.* **6**, 3836–3843 (2004).
20. Turekian, V. C., Macko, S. A. & Keene, W. C. Concentrations, isotopic compositions, and sources of size-resolved, particulate organic carbon and oxalate in near-surface marine air at Bermuda during spring. *J. Geophys. Res.* **108** (D5), 4157 (2003).
21. Nissenon, P., Knox, C. J. H., Finlayson-Pitts, B. J., Philips, L. F. & Dabdub, D. Enhanced photolysis in aerosols: evidence for important surface effects. *Phys. Chem. Chem. Phys.* **8**, 4700–4710 (2006).
22. Richards, N. K. *et al.* Nitrate ion photolysis in thin water films in the presence of bromide ions. *J. Phys. Chem. A* **115**, 5810–5821 (2011).
23. Zhou, X. *et al.* Snowpack photochemical production of HONO: a major source of OH in the Arctic boundary layer in springtime. *Geophys. Res. Lett.* **28**, 4087–4090 (2001).
24. Val Martin, M., Honrath, R. E., Owen, R. C. & Li, Q. B. Seasonal variation of nitrogen oxides in the central North Atlantic lower free troposphere. *J. Geophys. Res.* **113**, D17307 (2008).
25. Helas, G. & Warneck, P. Background NO_x mixing ratios in air masses over the North Atlantic Ocean. *J. Geophys. Res.* **86** (C8), 7283–7290 (1981).
26. Li, S., Matthews, J. & Sinha, A. Atmospheric hydroxyl radical production from electronically excited NO_2 and H_2O . *Science* **319**, 1657–1660 (2008).
27. Carr, S., Heard, D. E. & Blitz, M. A. Comment on “Atmospheric hydroxyl radical production from electronically excited NO_2 and H_2O ”. *Science* **324**, 336b (2009).
28. Ye, C. *et al.* Comment on “Missing gas-phase source of HONO inferred from Zeppelin measurements in the troposphere”. *Science* **326**, 1657–1659 (2015).
29. Savarino, J. *et al.* Isotopic composition of atmospheric nitrate in a tropical marine boundary layer. *Proc. Natl Acad. Sci. USA* **110**, 17668–17673 (2013).
30. Browne, E. C. *et al.* Observations of total RONO_2 over the boreal forest: NO_x sinks and HNO_3 sources. *Atmos. Chem. Phys.* **13**, 4543–4562 (2013).

Acknowledgements This research is funded by National Science Foundation (NSF) grants (AGS-1216166, AGS-1215712, and AGS-1216743). We would like to acknowledge operational, technical and scientific support provided by NCAR's Earth Observing Laboratory, sponsored by the National Science Foundation. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF.

Author Contributions Ye, C. and Zhou, X. designed and performed the field and laboratory studies, interpreted the data and write the manuscript with inputs from all the co-authors; Cantrell, C. and Ye, C. performed model simulations.

Author Information The data are available in our project data archive (http://data.eol.ucar.edu/master_list/?project=SAS). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to X.Z. (xianliang.zhou@health.ny.gov).

METHODS

Field measurements. Measurements on board the C-130 aircraft during the NOMADSS field campaign included HONO, pNO₃, NO_x, O₃, BrO, IO, OH radicals, HO₂ radicals, RO₂ radicals, aerosol surface area densities (for particle diameter <1 µm), VOCs, photolysis frequencies, and other meteorology parameters. Extended Data Table 1 summarizes the instrumentation, time resolution, detection limit, accuracy and references^{31–40} for our measurements.

HONO was measured by two long-path absorption photometric (LPAP) systems based on the Griess–Saltzman reaction³¹. Briefly, ambient HONO was scrubbed by deionized water in a 10-turn glass coil sampler. The scrubbed nitrite was then derivatized with 5 mM sulfanilamide (SA) and 0.5 mM *N*-(1-naphthyl)-ethylene-diamine (NED) in 40 mM HCl, to form an azo dye within 5 min. The azo dye was detected by light absorbance at 540 nm using a fibre optic spectrometer (LEDSPC-4, World Precision Instruments) with a 1-m liquid waveguide capillary flow cell (World Precision Instruments). ‘Zero-HONO’ air was generated by pulling ambient air through a Na₂CO₃-coated denuder to remove HONO and was sampled by the systems periodically to establish measurement baselines. Interference from NO_x, peroxyacetyl nitrate (PAN) and particulate nitrite, if any, was corrected by subtracting the baseline from the ambient air signal. Owing to the low collection efficiency of these interfering species in the sampling coil and their low concentrations, the combined interference signal was estimated to be less than 10% of the total signal in the clean MBL. Potential interference from peroxyacetic acid (HO₂NO₂) was suppressed by heating the Teflon perfluoroalkoxy alkanes (PFA) sampling line to 50 °C with a residence time of 0.8 s. The HO₂NO₂ steady-state concentration in the MBL was estimated to be less than 1 p.p.t.v. at a temperature of 23–26 °C in both flights, and thus interference from HO₂NO₂ was negligible⁴¹. Overall, the instrument baseline in the clean MBL was stable and low, and clear and strong ambient air signals (approximately ten times the detection limit) were observed. The accuracy of HONO measurements was confirmed by comparison with limb-scanning differential optical absorption spectroscopy (DOAS)³⁶. The agreement between these two instruments was very good in wide power plant plumes, where HONO mixing ratios exceeded the lower detection limits of both instruments (Extended Data Fig. 3). HNO₃ and pNO₃ were quantitatively collected with a coil sampler and a fritted glass disk sampler, respectively. The collected nitrate in the two channels were reduced to nitrite by two cadmium (Cd) columns, and determined using two LPAP systems^{31,32}. Zero air was generated to establish measurement baselines: for HNO₃ by passing the ambient air through a NaCl-coated denuder to remove HNO₃, and for pNO₃ through a Teflon filter and a NaCl-coated denuder to remove aerosol particles and HNO₃. Potential interference from HONO, NO_x and PAN was corrected by subtracting the baselines from the ambient air signals.

Ozone measurements were unavailable on 8 July 2013 and OH radical measurements were unavailable on 5 July 2013 owing to instrument malfunction. Steady-state OH radical concentrations were calculated and used in the budget analysis when OH radical measurements were not available⁴². Most of the parameters were observed at similar values during both flights, indicating that both flights captured the primary features of the local chemical environment. The lack of OH and O₃ measurements on the different flights had a negligible impact on our analysis. Several spikes in NO_x and aerosol surface area density detected from ship exhaust were excluded from the analysis.

Seventy-two-hour back-trajectories were calculated for both flights (Extended Data Fig. 1) with the Lagrangian particle dispersion model FLEXPART⁴³, version 9.02 (<http://flexpart.eu>), using six-hourly meteorological analysis data of the Global Forecasting System of the National Centers for Environmental Prediction (http://nomads.ncep.noaa.gov/txt_descriptions/GFS_half_degree_doc.shtml), interlaced with 3-h forecasts (0:00 UTC, 3:00 UTC, 6:00 UTC, 9:00 UTC, 12:00 UTC, 15:00 UTC, 18:00 UTC and 21:00 UTC), at a horizontal resolution of 0.5°. Every 5 min during the research flight 10,000 particles were released at the then-current position of the NSF/NCAR C-130 and followed back in time for 72 h. A ‘particle’ here refers to an infinitesimally small parcel of air, which is only affected by three-dimensional transport, turbulence and convection, and does not have any removal processes (no deposition, washout, sedimentation, chemical losses). Centroids shown in the figures are based on an algorithm⁴⁴ that reduces the residence probability distribution resulting from the locations of the 10,000 particles into five probable locations at each time interval. **Particulate nitrate photolysis experiment.** One aerosol sample was collected using a Teflon filter (Sartorius, pore size 0.45 µm, diameter 47 mm) on-board the NSF/NCAR C-130 aircraft during every research flight from 30 min after take-off to 30 min before landing. The total sampling volume ranged from 1.1 m³ to 1.5 m³ depending on the flight length (ranging from 6 h to 8 h). The filter sample was wrapped in aluminium foil and stored in a refrigerator until use. The pNO₃ photolysis rate constant was determined using the filter sample in the laboratory.

The photochemical experiments were conducted using a cylindrical flow reactor (inner diameter 10 cm, depth 1.5 cm) with a quartz window on the top, at a

temperature of 21.0 ± 1.0 °C and a relative humidity of (50 ± 2)%. The filter sample was moved into the flow cell directly from the freezer for the photochemical experiment. Compressed air was purified by flowing through an activated charcoal and Purafil chemisorbent column (Purafil) to remove NO_x, VOCs, H₂S, SO₂, HNO₃ and HONO and was used as carrier gas. Gaseous products, HONO and NO₂, released during the experiment were flushed out of the reactor by the carrier gas, and were sampled by two coil samplers connected in series. The first 10-turn coil sampler scrubbed HONO with purified water at a collection efficiency of 100% (ref. 31), and the second 32-turn coil sampler was to scrub NO₂ with an acetic acid modified SA/NED solution at a collection efficiency of 60% (ref. 45). The scrubbed nitrite and NO₂ were converted to an azo dye with SA/NED reagents and analysed by two separate LPAP systems^{31,45}. The filter sample was exposed to the solar simulator radiation for 10 min; baselines were established for both HONO and NO₂ before and after the light exposure. Photochemical production rates of HONO and NO₂ were calculated from their time-integrated signals above the baselines over the period of light exposure. To correct for HONO and NO₂ production from photolysis of HNO₃ deposited on the flow reactor wall surface, a control experiment was conducted by irradiating the empty flow reactor. The control signals were subtracted from the sample exposure signals when calculating the production rates of HONO and NO₂ from pNO₃ photolysis. After 10 min of light exposure, nitrate in the filter sample was extracted with 15 ml 1% NH₄Cl buffer solution (pH = 8.5), reduced to nitrite by a Cd column and determined by LPAP. A 300-W Ceramic xenon arc lamp (Perkin Elmer, model PE300BUV) was used as a light source. The ultraviolet light below 290 nm and the heat-generating infrared light were removed by a Pyrex glass filter and a water filter, respectively. The parabolic light beam irradiated only a circular area of a 1-inch (2.54 cm) diameter in the centre of the flow reactor. The shape of the spectrum of the filtered light source is similar to that of the solar actinic spectrum in the MBL (Extended Data Fig. 4). The effective light intensity in the centre of the flow reactor under direct irradiation was measured to be about 3.5 times higher than that at tropical noon on the ground (solar elevation angle $\theta = 0^\circ$)^{1,18}, using a nitrate actinometer¹⁸. The production rates of HONO and NO₂ were normalized by the amount of pNO₃ exposed and the effective ultraviolet light intensity, to obtain the normalized photolysis rate constants of pNO₃, $J_{\text{pNO}_3}^{\text{N}}$ (ref. 5).

$$J_{\text{pNO}_3}^{\text{N}} = \frac{P_{\text{HONO}} + P_{\text{NO}_2}}{N_{\text{nitrate}}} \times \frac{J_{\text{nitrate}, 0^\circ}}{J_{\text{nitrate}}} \quad (1)$$

where P_{HONO} and P_{NO_2} are the production rates of HONO and NO₂, respectively; N_{nitrate} is the light-exposed pNO₃ amount determined in the extraction solution; J_{nitrate} is the photolysis rate constant of nitrate in the actinometer solution exposed to the experimental light source; and the ‘standard’ photolysis rate constant of aqueous nitrate ($J_{\text{nitrate}, 0^\circ} \approx 3.0 \times 10^{-7} \text{ s}^{-1}$) is assumed for typical tropical summer conditions on the ground (solar elevation angle $\theta = 0^\circ$)^{1,18}, corresponding to a gas-phase HNO₃ photolysis rate constant ($J_{\text{HNO}_3, 0^\circ}$) of $\sim 7.0 \times 10^{-7} \text{ s}^{-1}$. HONO is the major product from pNO₃ photolysis, with an average production ratio of HONO to NO₂ of 2.0. It should be noted that the $J_{\text{pNO}_3}^{\text{N}}$ value is calculated from the production rates of HONO and NO₂, and thus it may underestimate the photolysis rate constant of pNO₃ if other products, such as NO, are produced. However, NO was only a minor secondary product from the photolysis of HONO and NO₂, and accounted for $\sim 1\%$ of the total production of HONO and NO₂ in the flow reactor in a residence time of 8 s. Therefore, lack of NO measurement in this study would not affect the $J_{\text{pNO}_3}^{\text{N}}$ measurement. The overall uncertainty in the photolysis rate constant measurement is about 50%, taking into account the measurement uncertainties in production rates of HONO and NO₂, nitrate loading and effective ultraviolet light intensity.

It should be pointed out that the laboratory-determined $J_{\text{pNO}_3}^{\text{N}}$ represents merely an average photolysis rate constant of pNO₃ collected during the entire flight covering a wide geographic area, and thus does not reflect the possible temporal and spatial variability in the actual $J_{\text{pNO}_3}^{\text{N}}$ during the flight. In addition, the photochemical reactivity of bulk aerosol sample collected on the filter may be altered during sampling and handling, and thus the laboratory-determined $J_{\text{pNO}_3}^{\text{N}}$ using the bulk aerosol sample might be different from that under the ambient conditions. Nevertheless, the laboratory-determined $J_{\text{pNO}_3}^{\text{N}}$ is still the first and best estimate of the pNO₃ photolysis rate constant in the ambient atmosphere so far, and is in reasonable agreement with the values inferred from field observations (Fig. 2).

To evaluate the global recycling NO_x source from particulate nitrate photolysis, more aerosol samples need to be collected from different atmospheric environments all over the world, and better-designed experiments need to be conducted using these samples to more accurately determine the photolysis rates constants of particulate nitrate under conditions similar to those of ambient atmosphere.

Model simulations. The air masses encountered consisted mostly of aged air masses circulating within the boundary layer of the North Atlantic Ocean under a Bermuda high-pressure system for several days before reaching the measurement

locations (Extended Data Fig. 1). Within this relatively isolated air mass in the MBL, the cycling of $\text{pNO}_3\text{--HONO--NO}_x$ was occurring much more rapidly, of the order of hours during the day, than dry deposition, which was of the order of days. Therefore, the diurnal chemistry of $\text{pNO}_3\text{--HONO--NO}_x$ in the MBL can be simulated by a zero-dimensional box model.

Model simulations were conducted using the MCM v3.2 updated with the Jet Propulsion Laboratory's latest recommended kinetics^{10,11,46}, and constrained by the field-measured long-lived species, such as O_3 , VOCs and total NO_x . The model was initiated from 00:00 local time and was allowed to run for three diurnal cycles (72 h). The diurnal concentration profiles of short-lived species, including OH, HO_2 , HONO and NO_x , were affected by their initial concentrations only during the first diurnal cycle of simulation. The results presented here (Fig. 3, Extended Data Figs 2 and 5) were from the second diurnal cycle of the simulation run. To simulate the time-varying cycling of $\text{pNO}_3\text{--HONO--NO}_x$, the diurnal profiles of photolysis frequencies were also calculated by the MCM v3.2 and scaled by the measured values in the early afternoon. The time-dependent photolysis rate constant of particulate nitrate (J_{pNO_3}) at certain times of the day was calculated as follows:

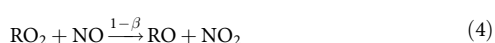
$$J_{\text{pNO}_3} = J_{\text{pNO}_3}^N \times \frac{J_{\text{HNO}_3}}{J_{\text{HNO}_3, 0^\circ}} \quad (2)$$

where J_{HNO_3} is the photolysis rate constant of gas-phase HNO_3 measured in the MBL. The ratio of pNO_3 to HNO_3 was set to 1, as observed in the MBL during the early afternoon.

The photolysis of pNO_3 is considered a source of both HONO and NO_x . To include all the possible HONO sources in the model, the following reactions are also considered: the gas-phase reactions of OH with NO, of excited NO_2 with H_2O (refs 26 and 27), and of $\text{HO}_2\text{--H}_2\text{O}$ with NO_2 (refs 16 and 28), and the heterogeneous reactions of NO_2 on sea-salt aerosol particles. Upper-limit HONO formation rates from reactions of excited NO_2 with H_2O and of $\text{HO}_2\text{--H}_2\text{O}$ with NO_2 are calculated using the latest recommendations^{27,28}. An upper-limit uptake coefficient of 10^{-4} was assumed for NO_2 uptake on sea-salt aerosol⁴⁷. HONO sinks in the model include its photolysis and the gas-phase reaction with OH radicals.

The photolysis of pNO_3 is a NO_x source mainly through HONO as an intermediate. Other NO_x sources, such as the photolysis of gaseous HNO_3 and the reaction of gaseous HNO_3 with OH, are negligible, and are not included here. NO_x sinks include the formation of bromine nitrate (BrONO_2), iodine nitrate (IONO_2) and organic nitrate (RONO_2) as well as HNO_3 through the reactions of NO_2 with OH and of NO with HO_2 . The uptake of halogen nitrate and organic nitrates on the sea-salt aerosols and following hydrolysis provide effective pathways for conversion of NO_x to particulate nitrate in addition to HNO_3 uptake. The photolysis lifetime of BrONO_2 in the gas phase was estimated to be ~ 12 min in the MBL, somewhat longer than the uptake and hydrolysis lifetime of ~ 6 min using an uptake coefficient of 0.8 on sea-salt aerosols⁴⁶. Therefore, two-thirds of BrONO_2 is converted to pNO_3 through hydrolysis on the aerosol particle at midday. The photolysis lifetime of IONO_2 in the gas phase was estimated to be ~ 5 min (ref. 48), comparable to its uptake and hydrolysis lifetime^{46,49}. Therefore, half of IONO_2 is converted to pNO_3 via hydrolysis on the aerosol particles at midday. BrO was measured at 1.5 p.p.t.v. by the DOAS instrument. The IO level was below the detection limit and was assumed to be at the typical MBL concentration^{49,50} of 0.5 p.p.t.v.

To evaluate the impact of VOCs on reactive nitrogen chemistry, explicit methane chemistry is considered in the model. Higher VOCs are lumped into a single species, RH. The concentration of RH is adjusted by the sum of the abundance of individually measured VOCs (R_i) scaled by the ratio of their rate coefficients for $\text{OH}+R_i\text{H}$ to that for $\text{OH}+\text{CH}_4$. The products of $\text{OH}+\text{RH}$ reactions lead to the formation of RO_2 radicals, carbonyls, peroxides, and organic nitrates. The yield of organic nitrate from the reaction of RO_2 radical with NO (reaction 3) is calculated based on the reaction rate of CH_3O_2 radical with NO ($k_{\text{CH}_3\text{O}_2+\text{NO}}$) and the effective branching ratio β .



The effective branching ratio is calculated:

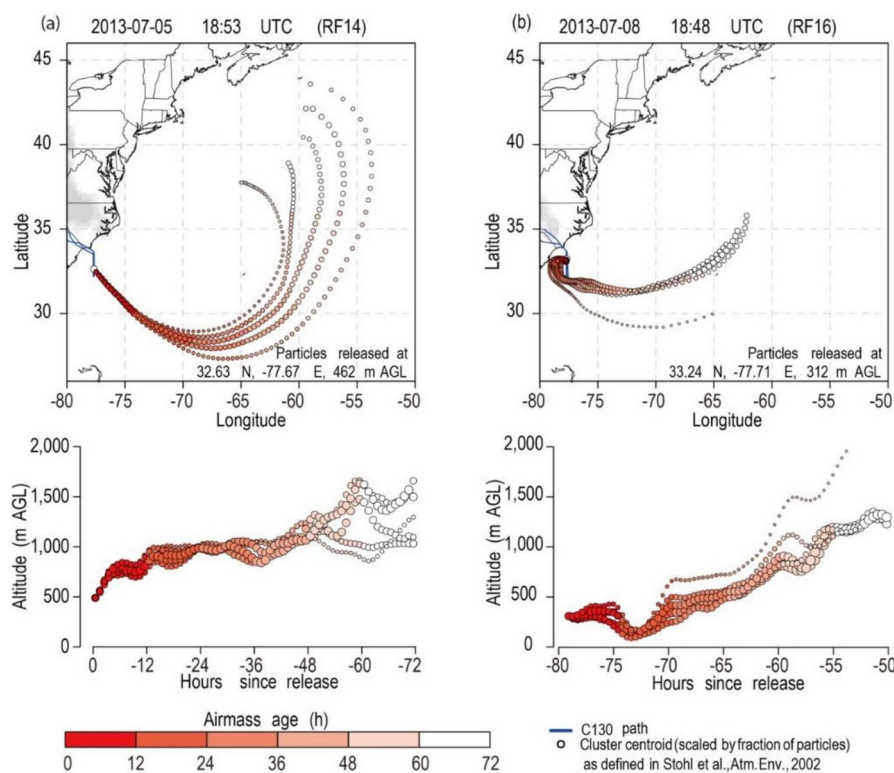
$$\beta = \frac{\sum k_i \beta_i [\text{R}_i\text{O}_2]}{k_{\text{CH}_3\text{O}_2+\text{NO}} [\text{CH}_3\text{O}_2]} \quad (5)$$

where $[\text{R}_i\text{O}_2]$ and $[\text{CH}_3\text{O}_2]$ are the concentrations of various R_iO_2 radicals formed from oxidation of various VOCs and methane, respectively; k_i and β_i are the recommended rate constants and branching ratios (β_i) for various RO_2 radicals, respectively^{30,51}. The effective branching ratio β is estimated to be $\sim 7\%$ for our model. The fate of RONO_2 in the MBL is not well known; hydrolysis has been assumed to be an

effective sink of organic nitrates in forested regions³⁰. In our model, an uptake coefficient of 10^{-2} and a 100% nitrate yield from RONO_2 uptake is assumed for organic nitrates on sea-salt aerosol.

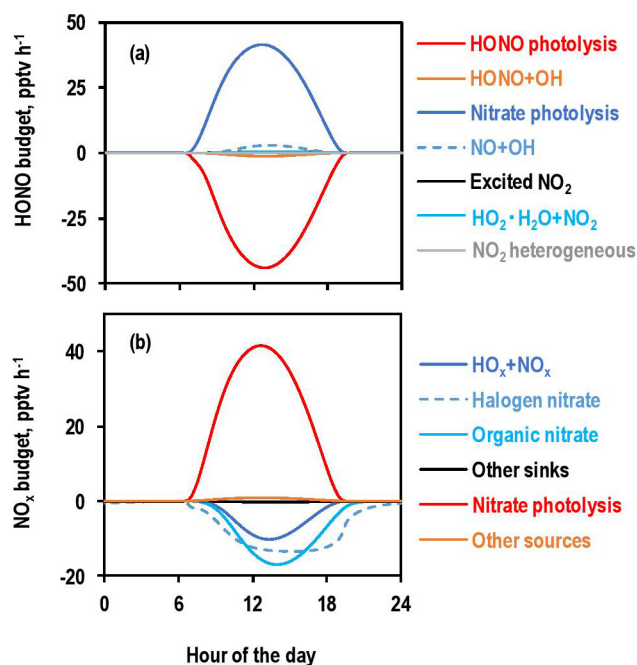
The model is initialized with typically measured parameters and is run at 15-min time steps for three diurnal cycles. Multiple runs are used to constrain the model uncertainty by error propagation from the errors of all considered parameters. Gaussian error propagation is applied for the uncertainty calculation. Overall, a model uncertainty of 28% (± 1 s.d.) is calculated. The reactive nitrogen species are found to be reproducible for multiple days without considering deposition and transport. Sensitivity studies with the model indicate that the calculated HONO concentrations are relatively insensitive ($\leq 2\%$) to the following assumed parameters: the RONO_2 branching ratio (varying from 1% to 7% to 14%), the uptake coefficients of RONO_2 and XONO_2 (changing up and down by a factor of two), the nitrate yield from RONO_2 hydrolysis (varying from 5% to 100%), the initial concentration of non-methane hydrocarbons (changing up and down by a factor of two), and the RO_2+NO reaction rate constant (changing up and down by a factor of two). Changing up and down by a factor of two in the initial concentrations of BrO or IO leads to a change of up to 4% in the calculated HONO. Finally, the sensitivity of the model output to the ratio of HNO_3 to pNO_3 was assessed by varying between 0.5 and 2.0, and the modelled HONO varied between 1.17 and 0.74 of the value for the ratio of unity. Despite the simplicity of this model approach, the good agreement between model calculations and field observations for key species (Fig. 3 and Extended Data Fig. 5) shows its feasibility and usefulness for assessing the chemistry of radicals and reactive nitrogen in the MBL.

- Zhang, N. *et al.* Measurements of ambient HONO concentrations and vertical HONO flux above a northern Michigan forest canopy. *Atmos. Chem. Phys.* **12**, 8285–8296 (2012).
- Huang, G., Zhou, X., Deng, G., Qiao, H. & Civerolo, K. Measurements of atmospheric nitrous acid and nitric acid. *Atmos. Environ.* **36**, 2225–2235 (2002).
- Ridley, B. *et al.* Florida thunderstorms: a faucet of reactive nitrogen to the upper troposphere. *J. Geophys. Res.* **109**, D17305 (2004).
- Mauldin, R. *et al.* South Pole Antarctica observations and modeling results: new insights on HO_x radical and sulfur chemistry. *Atmos. Environ.* **44**, 572–581 (2010).
- Hornbrook, R. S. *et al.* Measurements of tropospheric HO_2 and RO_2 by oxygen dilution modulation and chemical ionization mass spectrometry. *Atmos. Meas. Tech.* **4**, 735–756 (2011).
- Platt, U. & Stutz, J. *Differential Optical Absorption Spectroscopy: Principles and Applications* (Springer, 2008).
- Shetter, R. E., Cinquini, L., Lefer, B. L., Hall, S. R. & Madronich, S. Comparison of airborne measured and calculated spectral actinic flux and derived photolysis frequencies during the PEM Tropics B mission. *J. Geophys. Res.* **108** (D2), 8234 (2003).
- Flagan, R. C. Electrical mobility methods for sub-micrometer particle characterization. In *Aerosol Measurement: Principles, Techniques, and Applications* 3rd edn (eds Kulkarni, P. Baron, P. A. & Willeke, K.), 339–364 (John Wiley & Sons, 2011).
- de Gouw, J. & Warneke, C. Measurements of volatile organic compounds in the Earth's atmosphere using proton-transfer-reaction mass spectrometry. *Mass Spectrom. Rev.* **26**, 223–257 (2007).
- Hornbrook, R. S. *et al.* Observations of nonmethane organic compounds during ARCTAS—part 1: Biomass burning emissions and plume enhancements. *Atmos. Chem. Phys.* **11**, 11103–11130 (2011).
- Gierczak, T., Jimenez, E., Riffault, V., Burkholder, J. B. & Ravishankara, A. R. Thermal decomposition of HO_2NO_2 (peroxynitric acid, PNA): rate coefficient and determination of the enthalpy of formation. *J. Phys. Chem. A* **109**, 586–596 (2005).
- Cantrell, C. A. *et al.* Steady state free radical budgets and ozone photochemistry during TOPSE. *J. Geophys. Res.* **108** (D4), 8361 (2003).
- Stohl, A., Forster, C., Frank, A., Seibert, P. & Wotawa, G. The Lagrangian particle dispersion model FLEXPART version 6.2. *Atmos. Chem. Phys.* **5**, 2461–2474 (2005).
- Stohl, *et al.* A replacement for simple back trajectory calculations in the interpretation of atmospheric trace substance measurements. *Atmos. Environ.* **36**, 4635–4648 (2002).
- Zhang, N. *Distributions and Sources of HONO in the Rural Troposphere*. PhD thesis, <http://gradworks.umi.com/34/89/3489695.html> (State Univ. New York, 2011).
- Sander, S. *et al.* Chemical Kinetics and Photochemical Data for Use in Atmospheric Studies, Evaluation No. 17 JPL Publication 10-6, http://jpldataeval.jpl.nasa.gov/pdf/JPL09_16Nov09_Sander.pdf (Jet Propulsion Laboratory, 2011).
- Weis, D. D. & Ewing, G. E. The reaction of nitrogen dioxide with sea salt aerosol. *J. Phys. Chem. A* **103**, 4865–4873 (1999).
- Joseph, D. M., Ashworth, S. H. & Plane, J. M. C. On the photochemistry of IONO_2 : absorption cross section (240–370 nm) and photolysis product yields at 248 nm. *Phys. Chem. Chem. Phys.* **9**, 5599–5607 (2007).
- McFiggans, G. *et al.* A modeling study of iodine chemistry in the marine boundary layer. *J. Geophys. Res.* **105** (D11), 14371–14385 (2000).
- Dix, B. *et al.* Detection of iodine monoxide in the tropical free troposphere. *Proc. Natl Acad. Sci. USA* **110**, 2035–2040 (2013).
- Zhang, J., Dransfield, T. & Donahue, N. M. On the mechanism for nitrate formation via the peroxy radical + NO reaction. *J. Phys. Chem. A* **108**, 9082–9095 (2004).

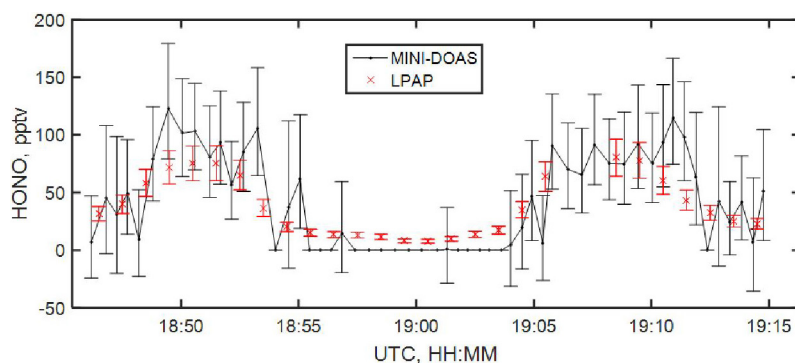


Extended Data Figure 1 | Typical back-trajectories of air masses in the MBL. a, 5 July 2013; b, 8 July 2013. The air mass circulated within the North Atlantic Ocean under a Bermuda high-pressure system for several days before reaching the measurement locations. On 8 July 2013, the air mass near the coast was also occasionally affected slightly by fresh

emissions from the southeast coast of the USA. See ref. 44 for definition of the cluster centroid. The C-130 flight tracks are shown in upper panels by the lines from 33° 39' N, 77° 42' W to 32° 11' N, 77° 41' W over the North Atlantic Ocean. The altitudes of cluster centroids in the lower panels are in metres above ground level (AGL).

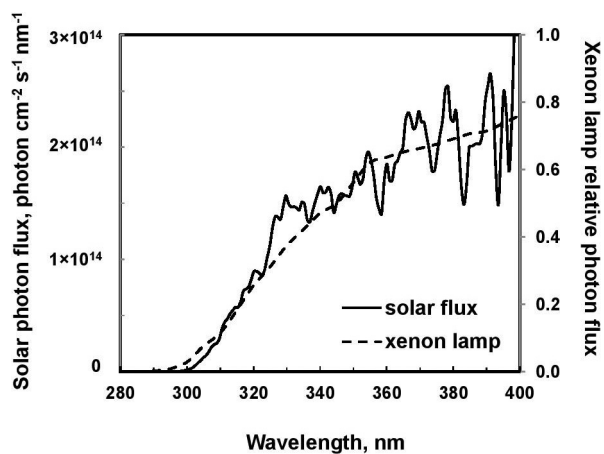


Extended Data Figure 2 | Typical diurnal HONO and NO_x budgets in the clean MBL calculated by the MCM v3.2. a, HONO budget; b, NO_x budget. 'HONO photolysis' and 'HONO+OH' represent HONO sinks contributed by HONO photolysis and the gas-phase reactions of HONO with the OH radicals. 'Nitrate photolysis', 'NO+OH', 'Excited NO₂', 'HO₂·H₂O+NO₂' and 'NO₂ heterogeneous' represent HONO sources contributed by pNO₃ photolysis, the gas-phase reactions of NO with OH, of excited NO₂ with H₂O and of HO₂·H₂O with NO₂, and the heterogeneous reactions of NO₂ on sea-salt aerosol particles, respectively. 'HO_x+NO_x' is the NO_x sink contributed by gas-phase reactions of NO₂ with OH and of NO with HO₂ with a branching ratio of 0.5% (ref. 46). 'Halogen nitrate' is the NO_x sink contributed by gas-phase reactions of NO₂ with primarily BrO and IO (ref. 29). 'Organic nitrate' is the NO_x sink contributed by reactions of RO₂ radicals with NO with an effective branching ratio of 7% (refs 30 and 51). 'Other sinks' represents other minor NO_x sinks, such as hydrolysis of N₂O₅. 'Nitrate photolysis' is the NO_x source contributed by pNO₃ photolysis and 'Other sources' represents other minor NO_x sources, such as photolysis of gaseous HNO₃.

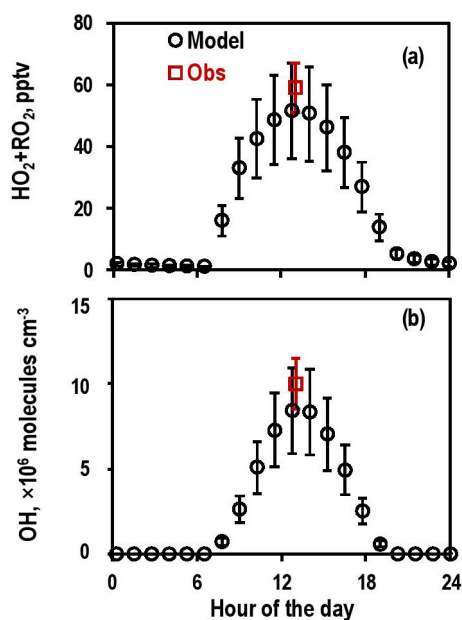


Extended Data Figure 3 | HONO observation comparison between the UCLA Mini-DOAS instrument and the LPAP instrument within two plumes during the research flight on 20 June 2013. The Mini-DOAS instrument is a limb-scanning Max-DOAS instrument, and the analysis incorporates the DOAS approach³⁶. The differential slant column densities of HONO were scaled to the differential slant column density analysis of HCHO, which were retrieved in the same spectral interval and multiplied by the *in situ* HCHO measurements, provided by the TOGA instrument⁴⁰ to derive HONO mixing ratios. The Mini-DOAS data was obtained along various elevation viewing angles from +45° to −10°. Because of the

position of the aircraft relative to the plumes, the plume geometry, and the HCHO scaling, the derived HONO mixing ratios do not depend on elevation viewing angle and this information was therefore omitted from the figure. Error bars (± 1 s.d.) accompany each DOAS measurement. DOAS measurements influenced by clouds or aircraft manoeuvres have been removed from the figure. For better visibility negative DOAS values, which were statistically indistinguishable from zero, were set to 0 p.p.t.v. and the error bar was removed. Red error bars show a 20% uncertainty (± 1 s.d.) for LPAP HONO results.



Extended Data Figure 4 | Comparison of the xenon arc lamp light spectrum with the solar actinic spectrum. The xenon light source was filtered by a 4-mm 7740 Pyrex filter, and the solar actinic flux was the 10-min measurement data at 360 m above sea level from 18:55 to 19:05 UTC during RF14.



Extended Data Figure 5 | Typical diurnal profiles of $\text{HO}_2 + \text{RO}_2$ and OH radicals in the clean MBL as observed and calculated by the MCM v3.2. **a**, $\text{HO}_2 + \text{RO}_2$ radicals; **b**, OH radicals. The calculated (Model) and measured (Obs) ratio of HO_2 radicals to RO_2 radicals (CH_3O_2 plus higher RO_2) is about 1. The error bars represent ± 1 s.d. of the model calculations and observations.

Extended Data Table 1 | Measurements from the NOMADSS study used in this analysis

Parameters	Instrument	Time Resolution	Detection Limit	Accuracy	Reference
HONO	LPAP	200 s	1 pptv	20%	(31)
pNO ₃	LPAP	360 s	5 pptv	30%	(31, 32)
HNO ₃	LPAP	20 min	5 pptv	30%	(31, 32)
NO	CI	1 s	10 pptv	10%	(33)
NO ₂	CI	1 s	20 pptv	15%	(33)
O ₃	CI	1 s	100 pptv	5%	(33)
OH	SICIMS	30 s	5×10^{-4}	30%	(34)
HO ₂ , RO ₂	SICIMS	60 s	30×10^{-6}	35%	(35)
BrO, IO	DOAS	60 s	1 pptv	20%	(36)
HONO	DOAS	60 s	~ 30 pptv	20%	(36)
Photolysis Frequencies	SAFS	1 s		18%	(37)
Surface area density	SMPS/UHSAS	65/1 s		20%	(38)
VOCs	PTRMS	15 s		20%	(39)
VOCs/organic nitrates	TOGA	120 s		20%	(40)

*In molecules per cm³.

Data are from refs 31–40. CI, three-channel chemiluminescence instrument; SICIMS, selected-ion chemical-ionization mass spectrometer; SAFS, scanning actinic flux spectroradiometer; SMPS, scanning mobility particle sizer; UHSAS, ultrahigh-sensitivity aerosol spectrometer; TOGA, trace organic gas analyser; PTRMS, proton-transfer reaction mass spectrometry.

Bubble accumulation and its role in the evolution of magma reservoirs in the upper crust

A. Parmigiani^{1,2}, S. Faroughi^{2,3}, C. Huber^{2,3}, O. Bachmann¹ & Y. Su²

Volcanic eruptions transfer huge amounts of gas to the atmosphere^{1,2}. In particular, the sulfur released during large silicic explosive eruptions can induce global cooling³. A fundamental goal in volcanology, therefore, is to assess the potential for eruption of the large volumes of crystal-poor, silicic magma that are stored at shallow depths in the crust, and to obtain theoretical bounds for the amount of volatiles that can be released during these eruptions. It is puzzling that highly evolved, crystal-poor silicic magmas are more likely to generate volcanic rocks than plutonic rocks^{4,5}. This observation suggests that such magmas are more prone to erupting than are their crystal-rich counterparts. Moreover, well studied examples of largely crystal-poor eruptions (for example, Katmai⁶, Taupo⁷ and Minoan⁸) often exhibit a release of sulfur that is 10 to 20 times higher than the amount of sulfur estimated to be stored in the melt. Here we argue that these two observations rest on how the magmatic volatile phase (MVP) behaves as it rises buoyantly in zoned magma reservoirs. By investigating the fluid dynamics that controls the transport of the MVP in crystal-rich and crystal-poor magmas, we show how the interplay between capillary stresses and the viscosity contrast between the MVP and the host melt results in a counterintuitive dynamics, whereby the MVP tends to migrate efficiently in crystal-rich parts of a magma reservoir and accumulate in crystal-poor regions. The accumulation of low-density bubbles of MVP in crystal-poor magmas has implications for the eruptive potential of such magmas^{9,10}, and is the likely source of the excess sulfur released during explosive eruptions.

Here, we use laboratory experiments and theoretical and numerical calculations to better understand the processes that control the dynamics of buoyant fluids in shallow magmatic systems. These fluids have a strong influence on the partitioning of volatile species and metals in magmas, as well as a deep impact on the scale and style of volcanic eruptions. Shallow magma reservoirs probably have stable, sharp transitions in crystallinity between crystal-rich and crystal-poor regions—an inference supported by geochemical, geological^{11,12} and thermal¹³ considerations. Crystal-poor caps form by progressive extraction of interstitial melts from mushy reservoirs^{14,15}. It is also commonly assumed, particularly for evolved arc magmas, that these incrementally built reservoirs contain exsolved volatiles^{1,2}. The MVP can be attributed to two sources: first, crystallization-driven exsolution ('second boiling') in the crystal mush; and second, periodic influx from degassing, underplating magma recharges.

In our model, we consider the buoyant migration of the MVP through an already formed mush–cap system (see Fig. 1 inset). In that context, the MVP is either produced directly by crystallization in the mush, or injected from below by periodic magma recharge. We argue for the existence of a process whereby the upward migration of the MVP is more efficient at high crystallinity, leading to an accumulation of MVP bubbles in shallower and less crystalline parts of the chamber. We do not include in our model the effect of concurrent crystallization of the mush, because crystallization in silicic mushes slows down

significantly when the magma temperature approaches the solidus¹⁶. In fact, if any crystallization does occur, it will tend to enhance MVP extraction by increasing confinement in the mush (see below).

In the crystal-poor cap, the MVP forms a bubble suspension in which the relative velocity of bubbles with respect to the melt is controlled by, first, hydrodynamic interactions between the melt and the bubbles¹⁷, and second, the effective buoyancy between the bubbles and melt. Coalescence is essentially precluded by the high melt viscosity and the low volume fraction, ψ , of MVP (less than 10–20 vol%). Combining a theoretical model with laboratory experiments (see Extended Data Fig. 1 and Methods)¹⁸, we find that bubble migration in viscous fluids slows down as the bubble volume fraction increases (solid line in Fig. 1). This effect is due to a reduction in buoyancy between a bubble and the suspension, and an increase in resistance to motion (drag from the enhanced return flow). These experiments also show that the development of bubble trains decreases the resistance of the bubbles to motion, but does not prevent this negative trend between bubble separation velocity and bubble volume fraction; the ascent of bubbles remains contingent on the downwelling of an equivalent volume of viscous melt.

In a crystal-rich environment, the solid volume fraction plays a key role in bubble migration. Viscous fingering (a heterogeneous displacement between two immiscible fluids) takes place when a non-wetting fluid (in our case, the MVP) is invading a porous medium filled with a more viscous wetting fluid¹⁹ (here, the silicate melt). Confinement by crystals, and the viscosity contrast between the MVP and the melt, enables vertically elongated fingering channels to grow and remain stable in the mush, as long as the flux of MVP is maintained (Fig. 2a and Supplementary Video 1).

Once established, the connected MVP network provides low-resistance pathways for MVP migration—that is, the rate of energy dissipation in the melt is reduced^{20,21}. We assume that the formation of these fingers does not deform the structure of the porous medium. The stress balance that controls fluid invasion involves a competition between viscous pressure drop, capillary stresses and friction between crystal grains²². Although the injection and migration of buoyant fluids can disturb the arrangement of a granular medium under low confining pressures^{22,23} and lead to capillary fracturing, we argue that the deformation of the crystal mush is negligible at a confining pressure greater than 1.5–2 kbars. This is for several reasons. First, crystals are angular with rough surfaces, and can form interpenetrative frameworks that drastically increase rigidity and friction²⁴. Second, under high normal stress, overcoming friction between crystals requires very large shear stresses at the pore scale. Third, in the mush, the pore pressure becomes significantly lower than the lithostatic pressure with depth; bubbles also need to overcome this pressure difference to deform the mush. Pore sizes of at most 0.1 μm should be needed for capillary forces to deform the mush²⁵. Fingering, rather than brittle or plastic deformation, is therefore the main MVP transport regime in these crystal mushes.

¹Institute of Geochemistry and Petrology, ETH Zurich, Zurich 8092, Switzerland. ²School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Georgia 30332, USA. ³School of Civil and Environmental Engineering, Georgia Institute of Technology, Georgia 30332, USA.

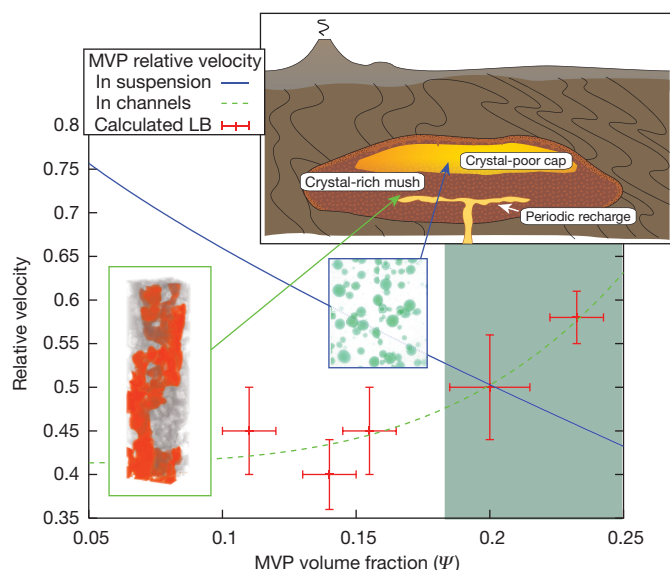


Figure 1 | Transport of a buoyant, non-wetting MVP in crystal-rich and crystal-poor magmas. Inset, representation of a crystal-rich mush zone underlying a crystal-poor cap. Main figure, the data show the relative velocity of bubbles in the cap (suspension; computed from equation (1); see Methods), and in the porous medium (crystallinity 50%; calculated by lattice Boltzmann (LB) simulations). The dashed line is a best-fit cubic power law for the relative velocity of the MVP in the porous medium. The error bars represent the range of steady-state velocities obtained when starting simulations with different initial conditions.

Assessing the processes that control the formation and efficiency of these MVP pathways in the porous medium requires a pore-scale approach (here, lattice Boltzmann simulations). Such an approach shows that crystal confinement promotes high-flux pathways for the MVP. Fingering becomes more stable as the MVP volume fraction, ψ , increases (above $\sim 10\%$) at a given crystal content (Fig. 1 and Extended Data Fig. 2), or as crystallinity increases at a given ψ (Fig. 2a). During the waning stage of degassing events, a significant fraction of the MVP (up to 10–15 vol%) can remain trapped in the mush by capillary and viscous forces along past flow pathways, leading to a residual saturation of MVP (Extended Data Fig. 2g). The residual MVP primes the mush for a more efficient outgassing during subsequent magma recharge events (see Methods for more information on the effect of fingering instabilities on MVP transport).

In magma reservoirs, the dynamics of MVP migration in the crystal-poor cap and the mush must be coupled, because one provides

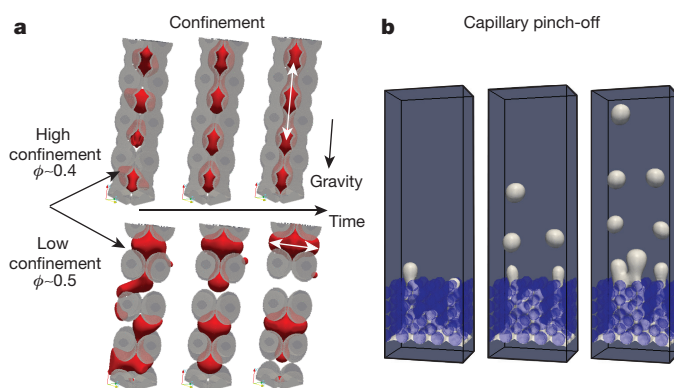


Figure 2 | Transport of a buoyant, non-wetting MVP in magmas with different crystal contents. a, The positive effect of crystal confinement on bubble coalescence and fingering formation. The MVP volume fraction is the same for both low- and high-confinement calculations. ϕ , porosity. b, Snapshots from a numerical simulation showing bubbles forming in a crystal-free environment as a result of the break-up of MVP fingers migrating through the underlying porous medium ('capillary pinch-off').

the flow boundary conditions for the other. At the mush–cap transition, spatial variations in crystallinity destabilize MVP fingers and lead to capillary pinch-off (fingers breaking into a stream of bubbles²¹; Fig. 2b). To resolve the force balance that controls the MVP migration at the mush–cap transition, we resort to lattice Boltzmann pore-scale multiphase flow calculations over a physical domain that extends a few centimetres on either side of that interface.

The set-up for the coupled mush–cap pore-scale calculations is based on the assumption that most of the MVP that transfers to the cap exsolves below the mush–cap domain that we model. This assumption is valid whether we consider second boiling in the mush or recharge to be the source of the MVP. Therefore, we implement an MVP flux boundary condition at the inlet of the model domain (underneath the porous layer). After injection, the MVP accumulates at the inlet, and fingering emerges naturally through the porous layer (Fig. 2b). The model set-up is consistent with the notion that fingering of MVP extracted from deeper in the system controls the volatile flux to the cap.

Our modelling results show that gas bubbles accumulate in the crystal-poor cap and that the accumulation efficiency grows with increasing viscosity contrast between the fluids (Fig. 3). The numerical calculations are limited to a range of viscosity ratios, λ , of $1/20 \leq \lambda = \nu_{nw}/\nu_w \leq 1$, where ν_{nw} is the kinematic viscosity of the

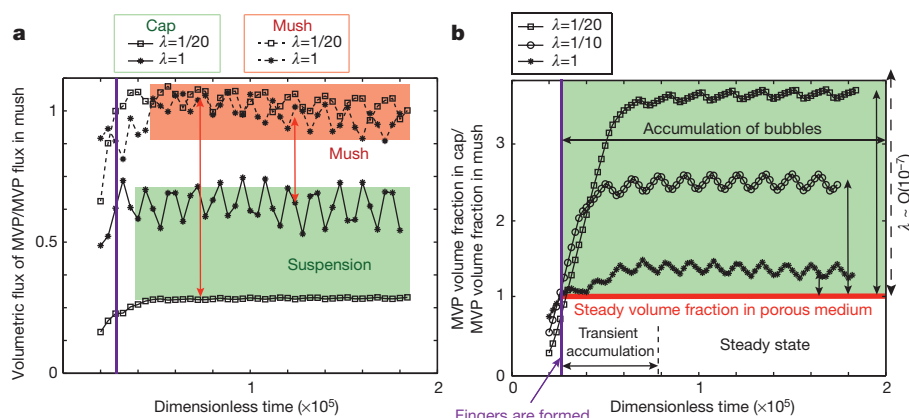


Figure 3 | MVP behaviour at the transition from crystal-rich to crystal-poor magma. a, b, Numerical calculations showing the change in transport regime of MVP from a confined medium (crystal-rich mush) to an unconfined horizon (crystal-poor cap; suspension). Panel a shows that the volumetric flux of MVP is greater in the mush than in the cap. Time

is normalized according to $t' = t \times \nu_{nw}/R^2$, where R is the radius of the capillary tube, ν_{nw} is the kinematic viscosity of MVP, and λ is ν_{nw}/ν_w . Panel b shows a comparison of the volume fraction of MVP in the mush (red line) and in the cap (black symbols) for different viscosity ratios between melt and volatiles.

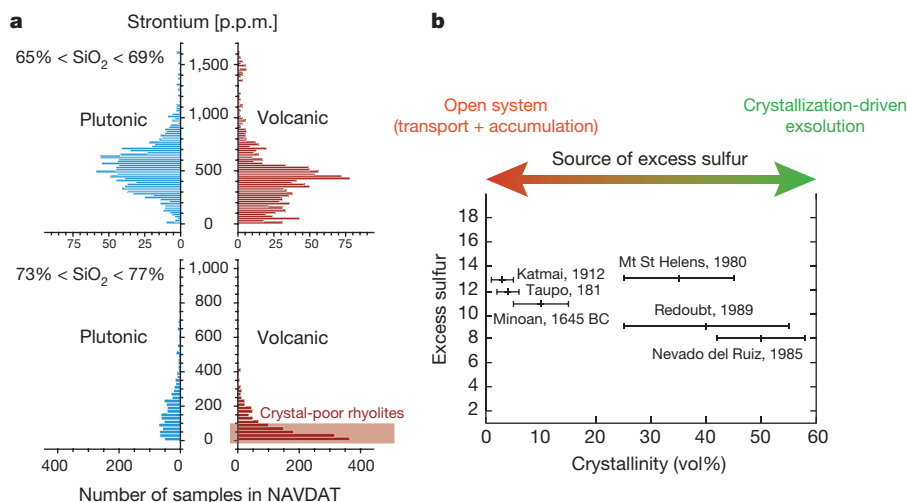


Figure 4 | Geological evidence for MVP accumulation in high-SiO₂, crystal-poor caps. a, Strontium content in whole-rock analyses of magmatic rocks from the NAVDAT database (<http://www.navdat.org>). While intermediate magmas (SiO₂ content between 65 wt% and 69 wt%) show a similar distribution of strontium in the plutonic and volcanic realms, the more silicic units (SiO₂ content between 73 wt% and 77 wt%)

show a significantly larger proportion of low-strontium, highly evolved, erupted magmas (<100 p.p.m. strontium) than do plutonic samples. **b**, Excess sulfur produced by selected explosive eruptions (predominantly crystal-poor: Katmai⁶, Taupo⁷ and Minoan⁸; crystal-rich: Mount St Helens³⁰, Redoubt³¹ and Nevado del Ruiz³²), and their sources of sulfur in excess of the amount dissolved in the melt.

non-wetting phase (MVP), and ν_w is the kinematic viscosity of the wetting phase (the silicate melt). For such a range, the MVP flux is up to five times higher in the mush than in the crystal-poor region (Fig. 3a), and the difference correlates positively with viscosity contrast. In silicic magma, λ is typically around 10^{-7} . Our calculations therefore provide a lower bound for the accumulation efficiency of MVP in crystal-poor horizons in magmatic systems (see Methods).

The contrasting dynamics of MVP migration in crystal-rich and crystal-poor environments is a consequence of the different processes that control the rate of energy dissipation. Once MVP pathways are established in the crystal-rich mush, dissipation is minimal because the more viscous melt is mostly passive (MVP flux in the mush is independent of the viscosity ratio, as shown in Fig. 3a). In suspensions, the absence of solid confinement promotes capillary break-up, and the buoyant ascent of bubbles is limited by a volumetrically equivalent return flow of silicate melt, which increases the viscous drag on other bubbles. The transition from one regime to the other is responsible for bubble accumulation in crystal-poor environments.

We note that convection is likely to occur in the crystal-poor cap, especially when subjected to the injection of MVP from the mush below. Convective stirring will affect the motion of bubbles in several ways. For example, it will lead to bubble entrainment in the convective cells, and homogenize the spatial distribution of bubbles in the cap, disrupting plumes forming near the vents at the mush–cap boundary. Overall, we find that convection promotes bubble accumulation (see Methods).

The accumulation of bubbles in crystal-poor caps has implications for the evolution of shallow magma reservoirs. For instance, upper crustal plutons are dominated by granodiorite/tonalite magma bodies¹¹, but deficient in granite *sensu stricto* (that is, they are deficient in evolved compositions with low concentrations of compatible elements⁴ such as strontium; Fig. 4a). In contrast, volcanic provinces can produce highly evolved (high-SiO₂, low-strontium), crystal-poor rhyolites that are volumetrically much more abundant than dacites (see, for example, the Yellowstone Province²⁶ and the Taupo Volcanic Zone²⁷). Bubble accumulation adds gravitational potential energy to crystal-poor rhyolitic caps; this favours the eruption of such liquids, rather than their stalling into the crust and forming granitic bodies⁵, and buffers the pressure drop during these eruptions¹⁰, allowing near-complete evacuation of the eruptible pockets of magma and the formation of very large volcanic units (as in, for example, supervolcanic eruptions

in Yellowstone²⁶, the Southern Rocky Mountain volcanic field²⁸ and the Taupo Volcanic Zone²⁷).

Last, but not least, bubble accumulation affects the volatile budget released during eruptions and pluton/ore formation. Magma chambers evacuated during, for example, the Katmai⁶, Taupo⁷ and Minoan⁸ eruptions are predominantly crystal-poor, and would not have undergone enough crystallization to yield their observed excess sulphur by second boiling. Hence, they require that sulfur-rich bubbles accumulate in the eruptible pods of magma (Fig. 4b). Crystal confinement can also have a significant role in the efficient devolatilization of magmas trapped in the crust and the release of metal-rich fluids that promotes the generation of the large porphyry copper systems associated with dying upper-crustal magma reservoirs^{20,29}.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 6 February 2015; accepted 26 January 2016.

Published online 13 April 2016.

- Wallace, P. J. Volcanic SO₂ emissions and the abundance and distribution of exsolved gas in magma bodies. *J. Volcanol. Geotherm. Res.* **108**, 85–106 (2001).
- Shinohara, H. Excess degassing from volcanoes and its role on eruptive and intrusive activity. *Rev. Geophys.* **46**, RG4005 (2008).
- Self, S. The effects and consequences of very large explosive volcanic eruptions. *Phil. Trans. R. Soc. A* **364**, 2073–2097 (2006).
- Halliday, A. N., Davidson, J. P., Hildreth, W. & Holden, P. Modeling the petrogenesis of high Rb/Sr silicic magmas. *Chem. Geol.* **92**, 107–114 (1991).
- Hildreth, W. Quaternary magmatism in the Cascades: geological perspectives. *US Dept Interior/US Geol. Surv. Prof. Pap.* 1744 (2007).
- Hildreth, W., Fierstein, J. Katmai volcanic cluster and the great eruption of 1912. *Geological Society of America Bulletin* **112**, 1594–1620 (2000).
- Wilson, C. J. N., Rogan, A. M., Smith, I. E. M., Northey, D. J., Nairn, I. A., Houghton, B. F. Caldera volcanoes of the Taupo Volcanic Zone, New Zealand. *J. Geophys. Res. Solid Earth* **89**, 8463–8484 (1984).
- Druitt, T. H., Costa, F., Deloule, E., Dungan, M. & Scaillet, B. Decadal to monthly timescales of magma transfer and reservoir growth at a caldera volcano. *Nature* **482**, 77–80 (2012).
- Blake, S. Volatile oversaturation during the evolution of silicic magma chambers as eruption trigger. *J. Geophys. Res.* **89**, 8237–8244 (1984).
- Huppert, H. E. & Woods, A. W. The role of volatiles in magma chamber dynamics. *Nature* **420**, 493–495 (2002).
- Bachl, C. A., Miller, C. F., Miller, J. S. & Faulds, J. E. Construction of a pluton: evidence from an exposed cross-section of the Searchlight pluton, Eldorado Mountains, Nevada. *Geol. Soc. Am. Bull.* **113**, 1213–1228 (2001).
- Heise, W., Caldwell, T. G., Bibby, H. M. & Bennie, S. L. Three-dimensional electrical resistivity image of magma beneath an active continental rift, Taupo Volcanic Zone, New Zealand. *Geophys. Res. Lett.* **37**, L10301 (2010).

13. Cooper, K. M. & Kent, A. J. R. Rapid remobilization of magmatic crystals kept in cold storage. *Nature* **506**, 480–483 (2014).
14. Bachmann, O. & Bergantz, G. W. On the origin of crystal-poor rhyolites: extracted from batholithic crystal mushes. *J. Petrol.* **45**, 1565–1582 (2004).
15. Hildreth, W. S. Volcanological perspectives on Long Valley, Mammoth Mountain, and Mono Craters: several contiguous but discrete systems. *J. Volcanol. Geotherm. Res.* **136**, 169–198 (2004).
16. Huber, C., Bachmann, O. & Manga, M. Homogenization processes in silicic magma chambers by stirring and mushification (latent heat buffering). *Earth Planet. Sci. Lett.* **283**, 38–47 (2009).
17. Faroughi, S. A. & Huber, C. A generalized equation for rheology of emulsions and suspensions of deformable particles subjected to simple shear at low Reynolds number. *Rheol. Acta* **54**, 85–108 (2015).
18. Faroughi, S. A. & Huber, C. Unifying the settling velocity in suspensions and emulsions of non-deformable particles. *Geophys. Res. Lett.* **42**, 53–59 (2015).
19. Lenormand, R., Touboul, E. & Zarcone, C. Numerical models and experiments on immiscible displacements in porous media. *J. Fluid Mech.* **189**, 165–187 (1988).
20. Huber, C., Bachmann, O., Vigneresse, J.-L., Dufek, J. & Parmigiani, A. A physical model for metal extraction and transport in shallow magmatic systems. *Geochem. Geophys. Geosys.* **13**, Q08003 (2012).
21. Parmigiani, A., Huber, C., Chopard, B. & Bachmann, O. Pore-scale mass and reactant transport in multiphase porous media flows. *J. Fluid Mech.* **686**, 40–76 (2011).
22. Holtzman, R., Szulczewski, M. L. & Juanes, R. Capillary fracturing in granular media. *Phys. Rev. Lett.* **108**, 264504 (2012).
23. Oppenheimer, J., Rust, A. C., Cashman, K. V. & Sandnes, B. Gas migration regimes and outgassing in particle-rich suspensions. *Front. Phys.* **3**, 60 (2015).
24. Philpotts, A. R., Shi, J. & Brustman, C. Role of plagioclase crystal chains in the differentiation of partly crystallized basaltic magma. *Nature* **395**, 343–346 (1998).
25. Jain, A. K. & Juanes, R. Preferential mode of gas invasion in sediments: grain-scale mechanistic model of coupled multiphase fluid flow and sediment mechanics. *J. Geophys. Res.* **114**, B08101 (2009).
26. Christiansen, R. L. The quaternary and Pliocene Yellowstone plateau volcanic field of Wyoming, Idaho, and Montana. *U.S. Geol. Surv. Prof. Pap.* 729-G (2001).
27. Graham, I. J., Cole, J. W., Briggs, R. M., Gamble, J. A. & Smith, I. E. M. Petrology and petrogenesis of volcanic rocks from the Taupo Volcanic Zone: a review. *J. Volcanol. Geotherm. Res.* **68**, 59–87 (1995).
28. Lipman, P. W. & Bachmann, O. Ignimbrites to batholiths: integrating perspectives from geological, geophysical, and geochronological data. *Geosphere* **11**, 705–743 (2015).
29. Sillitoe, R. H. Porphyry copper systems. *Econ. Geol.* **105**, 3–41 (2010).
30. Devine, J. D., Sigurdsson, H., Davis, A. N. & Self, S. Estimates of sulfur and chlorine yield to the atmosphere from volcanic eruptions and potential climatic effects. *J. Geophys. Res.* **89**, 6309–6325 (1984).
31. Gerlach, T. M., Westrich, H. R., Casadevall, T. J. & Finnegan, D. L. Vapor saturation and accumulation in magmas of the 1989–1990 eruption of Redoubt Volcano, Alaska. *J. Volcanol. Geotherm. Res.* **62**, 317–337 (1994).
32. Sigurdsson, H., Carey, S., Palais, J. M. & Devine, J. Pre-eruption compositional gradients and mixing of andesite and dacite magma erupted from Nevado del Ruiz, Colombia in 1985. *J. Volcanol. Geotherm. Res.* **41**, 127–151 (1990).

Supplementary Information is available in the online version of the paper.

Acknowledgements Discussion of an early version of the paper with A. Burgisser, P. W. Lipman, O. Malaspinas, M. Lupi and W. Degruyter helped us to clarify some concepts. We also thank O. Malaspinas and the rest of the Palabos team, as well as M. L. Porter for discussing how to implement lattice Boltzmann algorithms. We thank J. Bourquin for help with redrafting several figures. A.P. and O.B. acknowledge support from the Swiss National Science Foundation (Ambizione grant no. 154854 to A.P., and project no. 200021-103441 to O.B.). S.F., C.H. and Y.S. acknowledge funding from a National Science Foundation CAREER grant (1454821; awarded to C.H.). This work was also supported by grants from the Swiss National Supercomputing Centre (CSCS) under projects s479 and s597, and the Euler Supercomputer from ETHZ.

Author Contributions C.H., O.B. and A.P. conceived the research. C.H. and, to a lesser extent, A.P. developed the physical model. A.P. performed the numerical modelling and analysed the results. S.F. developed the laboratory experiments and theoretical model for the transport of volatiles in crystal-poor magmas. Y.S. led the discussion on excess sulfur. C.H., O.B. and A.P. all wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.P. (andrea.parmigiani@erdw.ethz.ch).

METHODS

Geological observations. Magma differentiation. Modelling of crystal fractionation in magmas predicts that the content of compatible trace elements such as strontium (highly enriched in plagioclase, the dominant mineral phase in the mid to upper continental crust) drops in the residual melt and should correlate with the SiO_2 content⁴. In geochemical datasets, such as the North American Volcanic and Intrusive Rock Database (NAVDAT; <http://www.navdat.org>), the distribution of strontium for a given SiO_2 range between volcanic and plutonic suites presents a striking conundrum. For dacitic/granodioritic compositions (intermediate SiO_2 content), the frequency of rock samples with similar strontium contents is nearly identical between volcanic and plutonic rocks; for high- SiO_2 magmas, there is a clear trend towards many more low-strontium compositions in volcanic units than in plutonic units. Clearly, geochemical datasets such as NAVDAT may carry sampling biases, but we argue that the number of samples considered (1,989 volcanic rock samples with 65–69% SiO_2 ; 2,355 volcanic rock samples with 73–77% SiO_2 ; 2,018 plutonic rock samples with 65–69% SiO_2 ; and 1,647 plutonic rock samples with 73–77% SiO_2 ; see Fig. 4a) and the prominence of the low-strontium mode in high- SiO_2 rhyolites imply that magmas depleted in strontium are more commonly represented in the eruptive than in the intrusive record. We assume that these low-strontium magmas are formed by interstitial/residual melt extraction from dacitic crystal-rich mushes, and segregate into liquid-dominated caps^{15,16,33,34} that are consequently more prone to erupt than to stall in the crust.

Volatile budget of volcanic eruptions. Estimates of volatile mass balance in magmas stored in the crust are generally based on the volatile content dissolved in melt inclusions³⁵. However, melt inclusion data are limited by several factors, including: (1) leakage of volatiles (loss through volume diffusion, cracks or cleavage surfaces); and (2) an inability to record volatiles trapped in hidden reservoirs (exsolved bubbles or sulfide phases). The excess sulfur paradox is a clear consequence of such limitations. Sulfur degassing from the melt during magma ascent in the conduit does not contribute to the excess sulfur because it is typically accounted for in the petrologic estimate from melt inclusions². The processes that govern the unbalanced sulfur budget are the build-up of an abundant sulfur-rich MVP produced by crystallization-driven exsolution, the transport and accumulation of sulfur-rich MVP from deeper untapped portions of the magmatic system, and, in some cases, the breakdown of sulfides or anhydrite before an eruption^{1,2,36,37}. In the context of the large excesses of sulfur released during the explosive eruption of crystal-poor rhyolitic caps, the contributions to the total sulfur mass budget from crystallization-driven exsolution and sulfide/anhydrite breakdown are bound to be tenuous, and call for an efficient migration and accumulation of MVP exsolved deeper down (by second boiling in the crystal-rich mushy roots of the magmatic system; see Fig. 4b).

MVP migration in the crystal-rich mush. At high crystallinity, buoyant bubbles are likely to deform along the direction perpendicular to gravity and, therefore, experience a significant hydrostatic pressure drop. Once this pressure drop is high enough to invade a pore throat, drainage is initiated and the blobs of MVP migrate vertically. The formation of anisotropic MVP clusters along the direction of gravity requires a confinement from the crystal phases to work against interfacial tension. Interfacial tension will tend to make bubbles spherical, whereas gravity will provoke the horizontal expansion of bubbles when their ascent is obstructed. Crystal confinement is therefore key to the development and stability of mobile MVP fingers in a crystal mush. In Extended Data Fig. 2, we show snapshots (Extended Data Fig. 2a–c show initial conditions and Extended Data Fig. 2d–f show the steady-state MVP distribution) for three numerical calculations conducted using a multiphase lattice Boltzmann model (that is, the interparticle potential Shan–Chen method^{21,38–40}). We implemented this model using the open-source Palabos library (<http://www.palabos.org>) and ran the simulations on the supercomputer clusters at Georgia Tech, the Swiss Federal Institute of Technology Zurich (ETHZ) and CSCS-Switzerland (the lattice Boltzmann method and code validations are described in detail below). In these calculations, the pore volume fraction of MVP and the size of pore throats are identical but with different crystallinities (crystallinity $(1-\phi)$ increases to the right). We see that increasing the spatial confinement (crystallinity) leads to enhanced coalescence and the formation of stable fingering. The run at highest crystallinity (Extended Data Fig. 2c, f) maximizes the vertical bubble pressure drop and is the only one that leads to the formation of a continuous fingering feature across the domain. In the other calculations, bubbles are mechanically trapped by capillary and viscous forces. A high MVP volume fraction and high crystallinity favour the formation and stability of viscous fingers, because they prevent the growth of Rayleigh–Plateau instabilities^{41–43} that are responsible for the break-up of fingering.

In essence, fingering pathways, once established, require little displacement of the viscous melt in the porous medium and reduce therefore the rate of energy dissipation in the melt. This results in an increase of MVP discharge in the mush. In Extended Data Fig. 2g, we report the results of a set of calculations (78 in total)

conducted with the same porous medium geometry (porosity 0.4) but a varying initial spatial distribution and volume fraction of MVP. The porous medium is made up of spherical pores, each connected to six cylindrical throats along each dimension (in three dimensions). Throat radii are randomly generated to introduce a random distribution of capillary entry pressure for the MVP invasion in neighbour pores. The calculations at a given MVP pore volume fraction are repeated with a different initial distribution of MVP in the pore space.

These calculations clearly show that the MVP discharge through the porous medium increases with the MVP pore volume fraction. At low MVP volume fraction (red region), MVP remains distributed as discrete bubbles trapped in the medium because of capillary and viscous forces. The discharge is negligible. At intermediate MVP pore volume fraction (green region), coalescence becomes important and makes the formation of percolating fingering pathways of MVP possible, which leads to a sharp increase in MVP discharge. However, in this region, the connectivity of fingering pathways depends on the initial distribution of MVP. In this regime, the runs that do not yield an efficient MVP discharge often display intermittent formation and destruction of fingering pathways, leading to successive periods of short-lived efficient transport and periods of capillary and viscous trapping of MVP bubbles. Above a certain pore volume fraction (blue region), the MVP always forms and sustains percolating fingering pathways and the MVP migration rate is fast.

Bubble suspension dynamics. The rising velocity of an isolated bubble through an infinite stagnant fluid can be described by the law derived by Hadamard and Rybczynski (reviewed in refs 44 and 45). However, when it comes to finding the velocity of a single bubble rising inside a cloud of bubbles, the dynamics becomes more complex because bubbles interact hydrodynamically with each other and with the ambient melt. For example, at low Reynolds number, the rising velocity of a trailing bubble aligned with another (lead) bubble along their direction of motion is greater than that of an individual bubble (the Smoluchowski effect; see ref. 46), while misaligned bubbles experience a greater viscous drag because of the melt return flow. Recently, Faroughi and Huber¹⁸ characterized both local and non-local bubble interactions theoretically, and proposed a new hindrance function, $F(\Psi, \lambda)$, which represents the ratio of the migration velocity of a bubble in a suspension to that of the same bubble in a bubble-free melt. The relative velocity of bubbles in a suspension at low Reynolds number is controlled by the balance between buoyancy and viscous stresses. The presence of bubbles decreases the hydrostatic pressure by a factor $(1-\Psi)$, whereas the presence of a cloud of MVP bubbles dispersed in the melt affects the effective shear viscosity of the magma¹⁷. The general expression for the hindrance function is¹⁸:

$$F(\Psi, \lambda) = \frac{1 - \Psi}{f_N^{cd} f_I^{cd} f_f^{\mu f}}$$

where f_N^{cd} represents the drag caused by the return flow of melt, parameterized as:

$$f_N^{cd}(\Psi, \lambda) = \left[1 - \frac{\beta}{2} \left(\frac{\Psi}{\Psi_m} \right)^{\frac{1}{3}} \left(\frac{2 + 3\lambda}{1 + \lambda} \right) + \frac{\beta^3}{2} \left(\frac{\Psi}{\Psi_m} \right) \frac{\lambda}{1 + \lambda} \right]^{-1}$$

Ψ and Ψ_m are, respectively, the bubble volume fraction and the random close packing limit for spherical bubbles; $\beta = 0.45$, being a geometrical proportionality constant determined experimentally¹⁷; f_I^{cd} captures Smoluchowski's effect:

$$f_I^{cd}(\Psi, \lambda) = 1 - \left[\frac{1}{2} \left(\frac{2 + 3\lambda}{1 + \lambda} \right) - \frac{1}{2} \frac{\Psi}{\Psi_m} \right]$$

and $f_f^{\mu f}$ expresses the change in momentum diffusivity via:

$$f_f^{\mu f}(\Psi, \lambda) = \frac{\mu_\Psi}{\mu_f} = \left(\frac{\Psi_m - \Psi}{\Psi_m(1 - \Psi)} \right)^{-\left(\frac{\Psi_m}{1 - \Psi_m} \right) \left(\frac{1 + 2.5\lambda}{1 + \lambda} \right)}$$

Finally, under the assumption that bubbles are inviscid relative to the melt, we obtain the relative bubble velocity:

$$F(\Psi, \lambda \rightarrow 0) = \frac{U_{\text{sus}}}{U_t} = \left(\frac{1 - \Psi}{1 - \frac{\Psi}{2\Psi_m}} \right) \left[1 - 0.45 \left(\frac{\Psi}{\Psi_m} \right)^{\frac{1}{3}} \left(\frac{\Psi_m - \Psi}{\Psi_m(1 - \Psi)} \right)^{\frac{\Psi_m}{1 - \Psi_m}} \right] \quad (1)$$

where U_{sus} and U_t are, respectively, the bubble velocity in the suspension and its Stokes ascent velocity. Equation (1) is plotted against experimental data over a wide range of particle volume fractions in Extended Data Fig. 1. We carried out experimental studies of bubble migration by using water injected at the top of a tank filled with silicon oil (Extended Data Fig. 3). The localized and fixed injection

points at the top of the tank (water is denser than silicon oils, so buoyancy is reversed compared with the typical situation in a magma reservoir) mimic the localized point sources that will transfer the MVP from the mush to the cap. The experimental set-up allows local bubble plumes to form, where hydrodynamic interactions introduce a smaller penalty to bubble buoyant migration. It is expected that bubble plumes ('vents') will form out of the mush in heterogeneous magma bodies; it was therefore necessary to validate equation (1) against this set of experimental data for our MVP cap suspension model (see Extended Data Fig. 1 inset). Note the significant decrease in MVP flux as the bubble fraction increases in a suspension; this contrasts strongly with the results shown in the section 'MVP migration in the crystal-rich mush', where the MVP flux increases significantly with increasing volume fraction in a porous mush.

Bubble residence time in crystal-poor caps. Crystal-poor caps are prone to convect, especially when buoyant bubbles are fluxed in from below. Convective motion will affect the migration of buoyant bubbles⁴⁷. At low Reynolds numbers, the overall motion of bubbles can be decomposed as a vectorial sum between the imposed convective motion and the buoyant phase separation calculated above. The behaviour of bubbles is determined by the ratio of these two velocity components (sometimes parameterized as a Stokes number⁴⁷). For small (millimetre-size) bubbles in a silicic magma, one can assume that bubbles remain highly coupled to the convective flow motion, except when the flow decelerates in boundary layers next to the edges of the reservoir. Thus, we adapt the model derived by Martin and Nokes⁴⁸ and also used by Dufek and Bachmann³⁴ for crystal suspensions to calculate the residence time of bubbles in the convecting cap.

The main differences between our calculations and those presented in refs 34 and 48 are: (1) in our calculations, the segregating phase comprises buoyant bubbles with free-slip conditions at the interface between bubbles and melt; and (2) we use the hindrance function derived above to correct for the presence of other bubbles, which can significantly affect the buoyant bubbles' ability to migrate in magmas. The model we obtain for the mass (here volume) conservation of bubbles in the cap therefore reads:

$$\frac{\partial}{\partial t} \left(\frac{\Psi}{1-\Psi} \right) = -\frac{U_i}{H} \left(\frac{\Psi}{1-\Psi} \right) F(\Psi, \lambda) + \frac{q}{H} \quad (2)$$

where $F(\Psi, \lambda)$ is the hindrance function calculated from equation (1), H is the thickness of the crystal-poor layer, and q is the volumetric flux of MVP coming from the mush.

We first solve equation (2) with $q=0$, and retrieve a characteristic residence time for bubbles in a convecting magma. In Extended Data Fig. 4, we show the solution to this differential equation under magmatic conditions. We find that increasing the initial volume fraction of bubbles in a convecting magma has a positive impact on accumulation—that is, at a higher volume fraction, bubbles remain trapped in the convective motion longer because of the hindrance to phase separation. Moreover, the decay rate of the bubble fraction that remains suspended in the convecting magma no longer follows an exponential law^{18,48}, because of the nonlinear dependence of the MVP ascent velocity on the MVP volume fraction. We also calculate the residence time of bubbles with two arbitrary sizes over a wide range of dynamic shear viscosities of the melt, for dilute ($\Psi=0.01$) and high ($\Psi=0.3$) volume fractions (Extended Data Fig. 5a). We determine the residence time as the half-life of bubbles in the cap, $\Psi(t_{1/2}) = 0.5\Psi_0$ (see ref. 19).

Under steady-state conditions, equation (2) reduces to:

$$\left(\frac{\Psi}{1-\Psi} \right) F(\Psi, \lambda) = Q \quad (3)$$

where $Q = q/U_i$ is a dimensionless sourcing term. We solve this equation to find the volume fraction of bubbles that can accumulate in the convecting layer, Ψ_c . The equation is nonlinear because of the hindrance function, and can admit more than one root. The physically meaningful solution is plotted in Extended Data Fig. 5b, and shows that the accumulated MVP volume fraction increases monotonously with the influx of MVP from the mush. Interestingly, equation (3) does not admit a real solution for injection rates that are greater than 15% of the Stokes final velocity of a 2-mm-diameter bubble in an infinite pool of melt with a dynamic viscosity of 10^6 Pa s (Extended Data Fig. 5b). Because these injection rates are quite modest, we expect that accumulation of bubbles up to a few tens of per cent in crystal-poor layers in magma chambers is possible. At higher injection rates, accumulation is still possible and likely to occur. However, the lack of a steady solution to our simple convecting suspension model implies that the multiphase dynamics will probably depart from that of a convecting suspension. We hypothesize that, as the volume fraction of bubbles in the

crystal-poor cap increases, more complex processes may arise and lead, for example, to massive Rayleigh–Taylor overturns⁴⁷.

Dynamic similarities with magma chamber dynamics. We explain the accumulation of MVP in crystal-poor horizons of magma reservoirs by the formation of continuous MVP fingers in crystal-rich environments^{21,40} and their break-up at the crystallinity transition between crystal-rich and crystal-poor magmas. This break-up of MVP fingers results in a significant change in the viscous dissipation regime.

We investigate this scenario numerically using a rather simplified geometry. We model the complex geometry of the crystal mush at the pore scale as a capillary tube that opens in a crystal-free/solid-free environment (Extended Data Fig. 6a, b). This is a simple proxy for the more realistic mush–cap transition, but it captures its essential ingredients: the dynamics of two immiscible fluids through a change in spatial confinement, where the low viscosity fluid is non-wetting and buoyant. We justify this approximation with the finding²¹ that the transport of immiscible fluids in a porous medium becomes mostly similar to an annular flow once the percolating pathway for the non-wetting fluid is reached. In our numerical calculations, a constant influx of MVP and a fixed pressure for the melt are set at the bottom boundary (inlet), while the top boundary (outlet) absorbs the outfluxing MVP and maintains a fixed pressure for the melt. The sides are periodic boundaries.

The competition between viscous, buoyancy, capillary and inertial forces controls both MVP transport and the breaking of continuous MVP fingering at the crystalline transition between crystal-rich and crystal-poor environments (bubble pinch-off frequency and volume^{49–51}). Because this balance operates at the pore scale, we resort to pore-scale multiphase flow calculations to study the formation and destruction of fingering pathways in a heterogeneous medium. The force balance can be described with three dimensionless numbers, the Archimedes (Ar), Bond (Bo) and Reynolds (Re) numbers. Ar, Bo and Re represent, respectively, the ratio between buoyancy and viscous forces (equation (4)), the ratio between buoyancy and capillary forces (equation (5)) and the ratio between inertia and viscous forces (equation (6)):

$$Ar = \frac{\rho_m \Delta \rho g D^3}{\mu_m^2} \quad (4)$$

$$Bo = \frac{\Delta \rho g D^2}{\sigma} \quad (5)$$

$$Re = \frac{\rho_m u_d D}{\mu_m} \quad (6)$$

where $\Delta \rho = \rho_{mvp} - \rho_m$ (ρ_{mvp} and ρ_m are the densities of MVP and melt), g is the acceleration due to gravity, μ_m the dynamic viscosity of the melt, D the bubble diameter and u_d the MVP average pore velocity. A rough estimate of these dimensionless numbers in shallow and highly evolved magmatic systems leads to $Ar \ll 1$, $Re \ll 1$ and $Bo \approx 0.1–1$, we obtain a Bo of the order of approximately 0.1 and we force Re and Ar to be lower than unity. Therefore, our results can serve as good first-order estimates for MVP accumulation in crystal-poor environments. The numerical method described in the 'Lattice Boltzmann for two-phase fluid flows' section limits us to relatively small viscosity contrasts compared with those expected in magmatic systems. Once pathways of MVP are established in the mush, the melt plays a passive role and does not affect the ascent of the MVP. The same is not true for the suspension, where the viscosity of the melt controls the rate of energy dissipation; as such, we expect accumulation to become more efficient as the viscosity contrast between the wetting and the less viscous non-wetting fluid increases. We decided to use our numerical model to test whether bubbles are likely to accumulate under less optimal conditions, that is, when the viscosity contrast is $1/20 \leq \lambda \leq 1$. We found that bubbles accumulate in the crystal-poor region even when the two fluids share the same viscosity ($\lambda = 1$), and that the accumulation potential increases as the viscosity contrast becomes more pronounced (Fig. 3).

Lattice Boltzmann for two-phase fluid flows. The lattice Boltzmann method (LBM) solves a discretized version of the continuum Boltzmann equation^{52,53}. Based on statistical mechanics, the LBM focuses on the mechanical interaction of an ensemble average distribution of particles $f_i(\mathbf{x}, t)$, and retrieves mass and momentum conservation (Navier–Stokes) equations from the statistical moment of the Boltzmann equation.

The LBM has been extended to multicomponent (MC) immiscible fluid flows. Among others, the MC Shan–Chen (SC) model^{38,54} is often applied because of:

(1) its straightforward implementation; and (2) the numerical stability of the algorithm in complex geometries such as porous media. In this work, we use the SC model extended by ref. 39, which allows us to model immiscible fluids characterized by notable viscosity contrast. These improvements result from an explicit formulation of the forcing term acting on the particle distribution functions and the use of a multi-relaxation-time (MRT) collision procedure. Below, we describe the improved algorithm briefly; for more details, see refs 39 and 55. *Explicit forcing and MRT collision operator.* The explicit evolution rule for the particle distribution function $f_i^\alpha(\mathbf{x}, t)$ with a single-relaxation-time (SRT) collision operator, Ω_i^α , can be written as:

$$f_i^\alpha(\mathbf{x} + \mathbf{e}_b, t+1) - f_i^\alpha(\mathbf{x}, t) = -\frac{1}{\tau_\alpha} (f_i^\alpha(\mathbf{x}, t) - f_i^{\text{eq}, \alpha}(\mathbf{x}, t)) + f_i^{F, \alpha} \left(1 - \frac{1}{2\tau_\alpha} \right) \quad (7)$$

where τ_α is the relaxation time for fluid A and B ($\alpha = A, B$) and relates to the fluids viscosity; $f_i^{\text{eq}, \alpha}$ is the equilibrium distribution function; and $f_i^{F, \alpha}$ is the explicit forcing term⁵⁶.

The left-hand side of equation (7) is generally referred to as the streaming of f_i^α values from the lattice node \mathbf{x} to one of its neighbours $\mathbf{x} + \mathbf{e}_i$; the right-hand side (the collision operator Ω_i^α) describes the exchange of momentum between the colliding f_i values. In equation (7), \mathbf{e}_i are a set of velocity vectors connecting nearest neighbour nodes (the spatial discretization of the lattice). Here we use the D3Q19 lattice—a three-dimensional lattice in which each node is connected to 19 neighbours. Lattice velocities \mathbf{e}_i and weights w_i for a D3Q19 lattice can be found in ref. 57. The equilibrium distribution function and the explicit forcing term in equation (7) read respectively:

$$f_i^{\text{eq}, \alpha}(\mathbf{x}, t) = w_i \rho_\alpha \left(1 + \frac{\mathbf{e}_i \bullet \mathbf{u}^{\text{eq}}}{c_s^2} + \frac{(\mathbf{e}_i \bullet \mathbf{u}^{\text{eq}})^2}{2c_s^4} + \frac{(\mathbf{u}^{\text{eq}})^2}{2c_s^2} \right)$$

$$f_i^{F, \alpha}(\mathbf{x}, t) = \frac{\mathbf{F}_\alpha \bullet (\mathbf{e}_i - \mathbf{u}^{\text{eq}})}{\rho_\alpha c_s^2} f_i^{\text{eq}, \alpha}$$

Here, c_s is the lattice speed of sound and \mathbf{u}^{eq} is the fluid mixture velocity defined as:

$$\mathbf{u}^{\text{eq}} = \frac{\sum_\alpha \rho_\alpha \mathbf{u}_\alpha \omega_\alpha}{\sum_\alpha \rho_\alpha \omega_\alpha}$$

where $\omega_\alpha = 1/\tau_\alpha$ and the statistical moments ρ_α (density) and $\rho_\alpha \mathbf{u}_\alpha$ (momentum) are calculated respectively as:

$$\rho_\alpha = \sum_i f_i^\alpha; \quad \rho_\alpha \mathbf{u}_\alpha = \sum_i f_i^\alpha \mathbf{e}_i + \frac{1}{2} \mathbf{F}_\alpha$$

The forcing vectors \mathbf{F}_α contain several contributions, notably cohesion (particle-particle), adhesion (particle-wall) and bulk (for example, gravity and buoyancy) forces. The cohesion forces, responsible for the physical separation between immiscible components, are calculated as:

$$\mathbf{F}_\alpha^{\text{coh}}(\mathbf{x}, t) = -\rho_\alpha(\mathbf{x}, t) G^c \sum_i w_i \rho_\beta(\mathbf{x} + \mathbf{e}_i) \mathbf{e}_i$$

where α and β are the two complementary phases and G^c is a free parameter that is used to tune the interfacial tension between the two fluids. The magnitude of the repulsive force applied by fluid B on fluid A at the node \mathbf{x} (and vice versa) depends on the density gradient of fluid B (for example, $\nabla \rho_B = \sum_i w_i \rho_B(\mathbf{x} + \mathbf{e}_i)$). The evaluation of $\nabla \rho_\alpha$ is critical for the stability of the calculations. High-density gradients (thin fluid–fluid interfaces) require an extended neighbourhood to reach the required accuracy⁵⁸. However, a better evaluation of density gradients comes at the price of an increase in computational time (especially in three dimensions). See refs 55, 59 for a detailed description of how to include adhesive and bulk forces. Here, in order to keep the numerical performance acceptable, we calculate the density gradients using the nearest neighbours only.

In order to improve the stability and accuracy of the SRT SC algorithm described above, we use an MRT collision operator, $\Omega_\alpha^{\text{MRT}}$. Then, the linear collision operator is re-cast into the space of velocity moments $\mathbf{m} = \mathbf{M} \times \mathbf{f} = (\mathbf{m}_0, \mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_{17}, \mathbf{m}_{18})$ (where \mathbf{M} is the transformation matrix⁵⁷); next, the relaxation parameter of each moment is adjusted individually to improve numerical stability. $\Omega_\alpha^{\text{MRT}}$ can be written as:

$$\Omega_\alpha^{\text{MRT}} = -\mathbf{M}^{-1} \bullet \mathbf{S}^\alpha \times (\mathbf{m}^\alpha - \mathbf{m}^{\text{eq}, \alpha}) + \frac{1}{2} \mathbf{M}^{-1} \bullet \mathbf{S}^\alpha \bullet \mathbf{m}^{F, \alpha} + \mathbf{f}^{F, \alpha}$$

In this equation, \mathbf{S}^α are diagonal matrices where the 19 diagonal components represent the relaxation parameter for each moments of f_i^α . As suggested in ref. 57, we use:

$$\mathbf{S}^\alpha = \text{diag} \left(1, 1.19, 1.4, 1, 1.2, 1, 1.2, 1, 1.2, \frac{1}{\tau_\alpha}, 1.4, \frac{1}{\tau_\alpha}, 1.4, \frac{1}{\tau_\alpha}, \frac{1}{\tau_\alpha}, \frac{1}{\tau_\alpha}, 1.98, 1.98, 1.98 \right)$$

The 19 components of the vectors \mathbf{m}^α , $\mathbf{m}^{\text{eq}, \alpha}$ and $\mathbf{m}^{F, \alpha}$ can be calculated respectively as:

$$\mathbf{m}_i^\alpha = \sum_j \mathbf{M}_{ij} f_j^\alpha; \quad \mathbf{m}_i^{\text{eq}, \alpha} = \sum_j \mathbf{M}_{ij} f_j^{\text{eq}, \alpha}; \quad \mathbf{m}_i^{F, \alpha} = \sum_j \mathbf{M}_{ij} f_j^{F, \alpha}$$

The stability of the algorithm depends mainly on the choice of repulsion constant (G^c) and its correspondent value at solid wall nodes (G^{wall} , used to introduce wetting forces). Here we want to deal with a highly non-wetting MVP phase. The non-wetting behaviour of MVP affects its dynamics both in the porous medium (higher capillary entry pressures) and at the transition between crystal-rich and crystal-poor environments (pinch-off dynamics).

LB algorithm validation. In order to validate the MRT SC multicomponent algorithm that we use to model the capillary finger formation and the pinch-off dynamics (Figs. 2 and 3), we test our model with two benchmarks. The first test is an annular Poiseuille flow, where the non-wetting fluid A is located in the centre of the pipe such that $r < R_{\text{in}}$, and the wetting fluid B is placed in the outer ring such that $R_{\text{in}} \leq r \leq R_{\text{out}}$ (where R is the radius of the pipe flow). Both fluids are accelerated by the same bulk force F^b . For the case of the two-phase Poiseuille profile problem, an analytical solution exists:

$$u(r) = \frac{F^b}{2\nu_A \rho_A} (R_{\text{in}}^2 - r^2) + \frac{F^b}{2\nu_B \rho_B} (R_{\text{out}}^2 - R_{\text{in}}^2), 0 \leq |r| \leq R_{\text{in}} \quad (8)$$

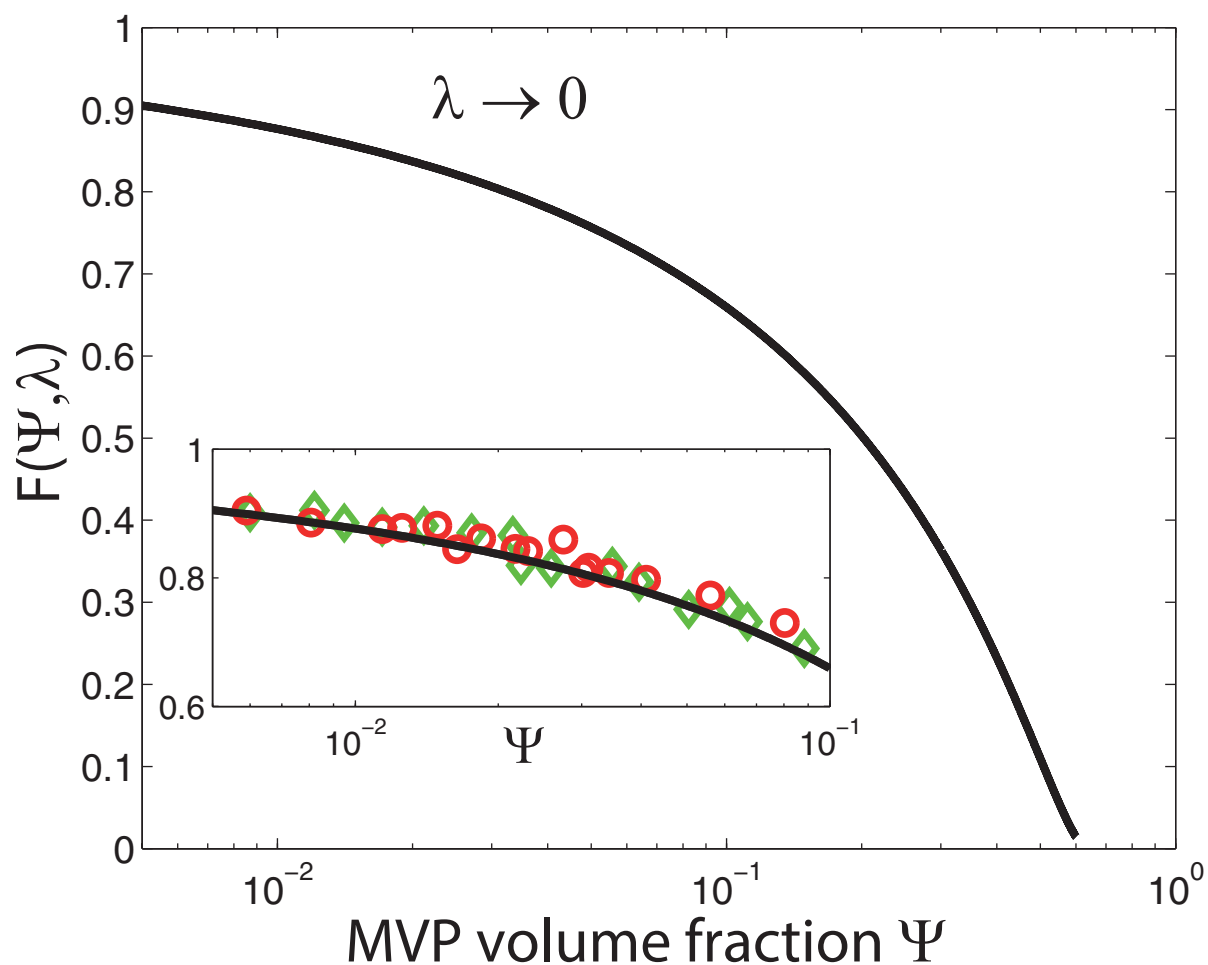
$$u(r) = \frac{F^b}{2\nu_B \rho_B} (R_{\text{out}}^2 - r^2), R_{\text{in}} \leq |r| \leq R_{\text{out}} \quad (9)$$

where ν_α is the kinematic viscosity of either fluid. In Extended Data Fig. 7a–c, we compare the analytical and numerical solutions for three different viscosity ratios ($\lambda = 1/5, 1/10$, or $1/20$).

The second validation test is a three-dimensional calculation of the equilibrium shape of a drop of fluid A embedded in fluid B and in contact with a flat solid surface. The goal of this validation is to reproduce the correct equilibrium (static) contact angle between the fluids and solid phases for different wetting properties (Extended Data Fig. 7d–f).

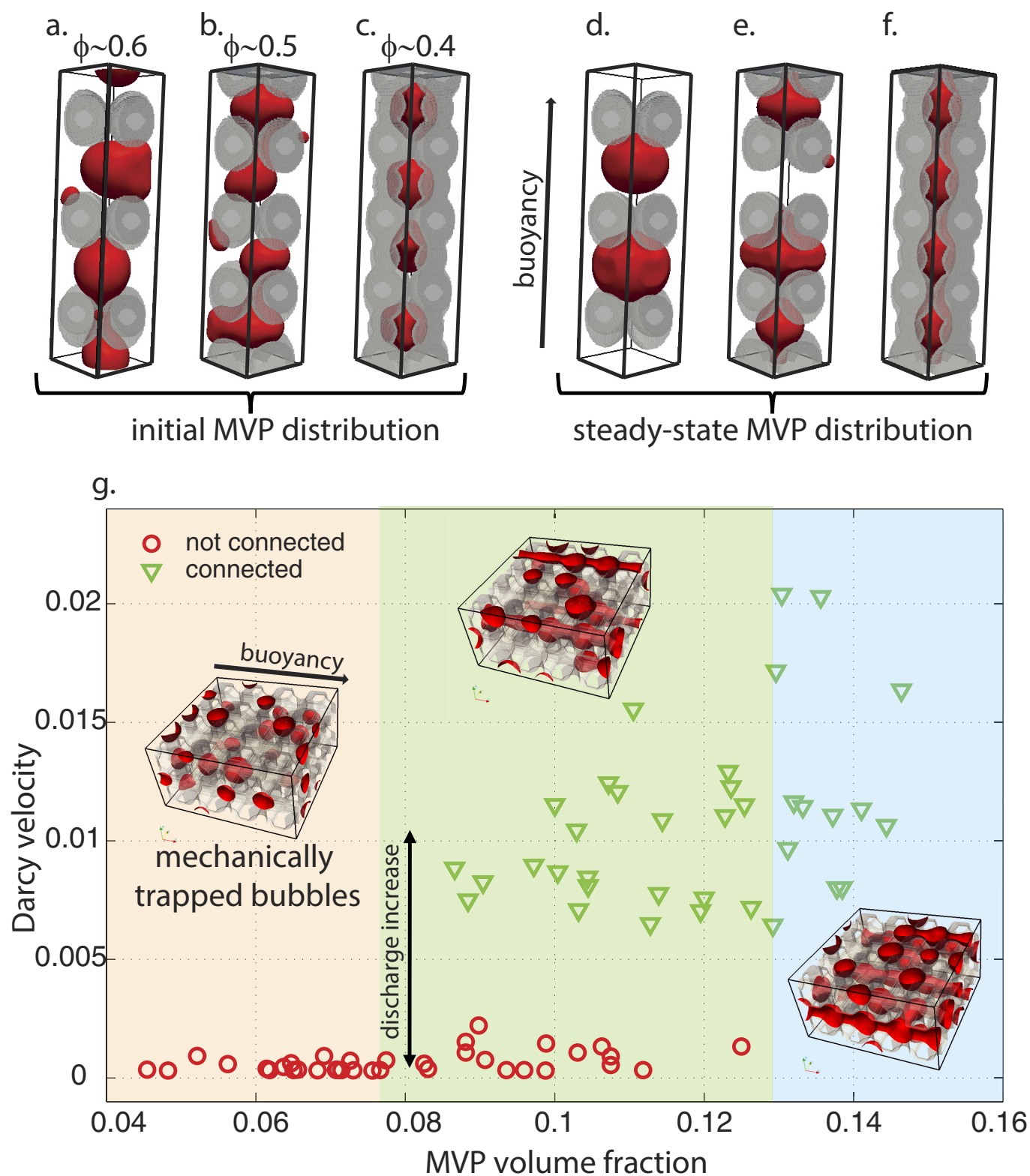
- Walker, B. J., Miller, C. F., Lowery, L. E., Wooden, J. L. & Miller, J. S. Geology and geochronology of the Spirit Mountain batholith, southern Nevada: implications for timescales and physical processes of batholith construction. *J. Volcanol. Geoth. Res.* **167**, 239–262 (2007).
- Dufek, J. & Bachmann, O. Quantum magmatism: magmatic compositional gaps generated by melt-crystal dynamics. *Geology* **38**, 687–690 (2010).
- Wallace, P. J. Volatiles in subduction zone magmas: concentration and fluxes based on melt inclusion and volcanic gas data. *J. Volcanol. Geotherm. Res.* **140**, 217–240 (2005).
- Gerlach, T. M., Westrich, H. R. & Symonds, R. B. in *Fire and Mud: Eruptions and Lahars of Mt. Pinatubo*, Philippine 415–433 (Univ. Washington Press, 1996).
- Costa, F., Scaillet, B. & Gourgand, A. Massive atmospheric sulfur loading of the AD 1600 Huaynaputina eruption and implications for petrological sulfur estimates. *Geophys. Res. Lett.* **30**, 1068 (2003).
- Shan, X. & Chen, H. Lattice Boltzmann model for simulation of flows with multiple phases and components. *Phys. Rev. E* **47**, 1815 (1993).
- Porter, M. L., Coon, E. T., Kang, Q., Moulton, J. D. & Carey, J. W. Multicomponent interparticle-potential lattice Boltzmann model for fluids with large viscosity ratios. *Phys. Rev. E* **86**, 036701 (2012).
- Huber, C., Parmigiani, A., Latt, J. & Dufek, J. Channelization of buoyant non-wetting fluids in saturated porous media. *Wat. Resour. Res.* **49**, 6371–6380 (2013).
- Gauglitz, P. A., St. Laurent, C. M. & Radke, C. J. Experimental determination of gas-bubble break-up in constricted cylindrical capillary. *Ind. Eng. Chem. Res.* **27**, 1282–1291 (1988).
- Guillot, P. & Colin, A. Stability of a jet in confined pressure-driven biphasic flows at low Reynolds number in various geometries. *Phys. Rev. E* **78**, 016307 (2008).
- Beresnev, I. A., Li, W. & Vigil, R. D. Condition for break-up of non-wetting fluid in sinusoidally constricted capillary channels. *Transp. Porous Media* **80**, 581–604 (2009).

44. Clift, R., Grace, J. R. & Weber, M. E. *Bubbles, Drops, and Particles*. (Courier Dover Publications, 1975).
45. Batchelor, G. K. *An Introduction to Fluid Dynamics*. (Cambridge Univ. Press, 1967).
46. Sonshine, R. M., Cox, R. G. & Brenner, H. The Stokes translation of a particle of arbitrary shape along the axis of a circular cylinder. *Appl. Sci. Res.* **16**, 273–300 (1966).
47. Ruprecht, P., Bergantz, G.W. & Dufek, J. Modeling of gas-driven magmatic overturn: tracking of phenocryst dispersal and gathering during magma mixing. *Geochem. Geophys. Geosys.* **9**, Q07017 (2008).
48. Martin, D. & Nokes, R. Crystal settling in a vigorously convecting magma chamber. *Nature* **332**, 534–536 (1988).
49. Higuera, F. J. Injection and coalescence of bubbles in a very viscous liquid. *J. Fluid Mech.* **530**, 369–378 (2005).
50. Gerlach, D., Alleborn, N., Buwa, V. & Durst, F. Numerical simulation of periodic bubble formation at a submerged orifice with constant gas flow rate. *Chem. Eng. Sci.* **62**, 2109–2125 (2007).
51. Quan, S. & Hua, J. Numerical studies of bubble necking in viscous liquids. *Phys. Rev. E* **77**, 066303 (2008).
52. He, X. & Luo, L. S. A priori derivation of the lattice Boltzmann equation. *Phys. Rev. E* **55**, R6333–R6336 (1997).
53. Shan, X., Yuan, X. F. & Chen, H. Kinetic theory representation of hydrodynamics: a way beyond the Navier-Stokes equation. *J. Fluid Mech.* **550**, 413–441 (2006).
54. Shan, X. & Doolen, G. D. Multicomponent Lattice Boltzmann model with interparticle interaction. *J. Stat. Phys.* **81**, 379 (1995).
55. Coon, E. T., Porter, M. L. & Kang, Q. Taxila LBM: a parallel, modular lattice Boltzmann framework for simulating pore-scale flow in porous media. *Comput. Geosci.* **18**, 17–27 (2014).
56. He, X., Chen, S. & Doolen, G. D. A novel thermal model for the lattice Boltzmann method in incompressible limit. *J. Comput. Phys.* **146**, 282–300 (1998).
57. D'Humieres, D., Ginzburg, I., Krafczyk, M., Lallemand, P. & Luo, L. S. Multiple-relaxation-time lattice Boltzmann models in three dimensions. *Phil. Trans. R. Soc. Lond. A* **360**, 437–451 (2002).
58. Sbragaglia, M. et al. Generalized lattice Boltzmann method with multirange pseudopotential. *Phys. Rev. E* **75**, 026702 (2007).
59. Huang, H., Thorne, D. T., Schaap, M. G. & Sukop, M. C. Proposed approximation for contact angles in the Shan-and-Chen type multicomponent multiphase lattice Boltzmann models. *Phys. Rev. E* **76**, 066701 (2007).



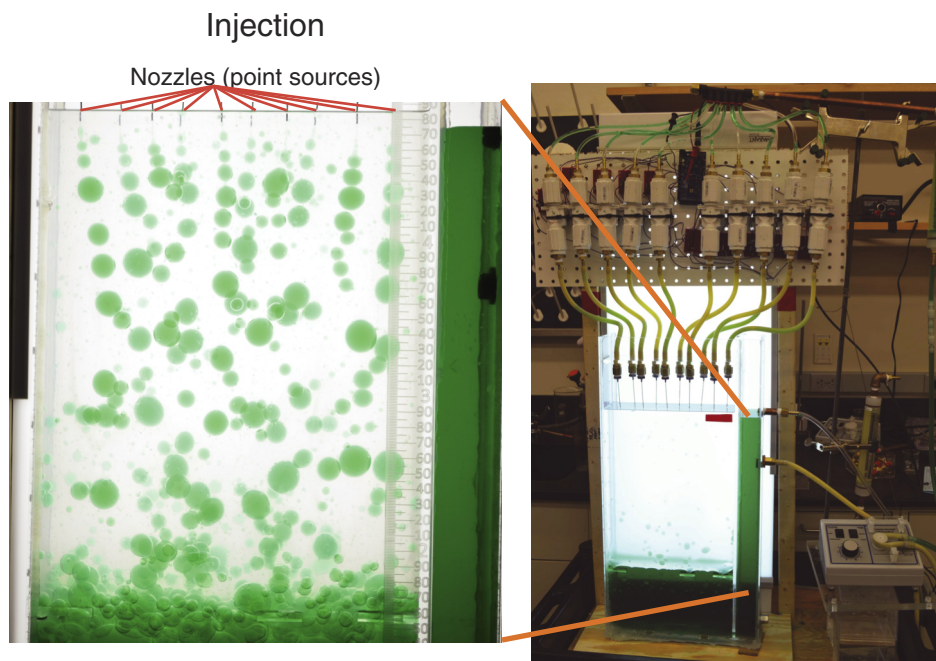
Extended Data Figure 1 | Hindrance function. The hindrance function, $F(\Psi, \lambda)$, defined by equation (1), for suspensions of MVP ($\lambda \rightarrow 0$) over a wide range of MVP volume fractions ($0 \leq \Psi \leq 0.6$). The inset shows the comparison of $F(\Psi, \lambda \rightarrow 0)$ with experimental data up to MVP volume

fractions of 10%. Experimental data are taken from ref. 18, where the method of continuous injection is used, injecting the dispersed phase (water) into the highly viscous ambient phase (silicone oil, resulting in $\lambda = O(10^{-4})$).

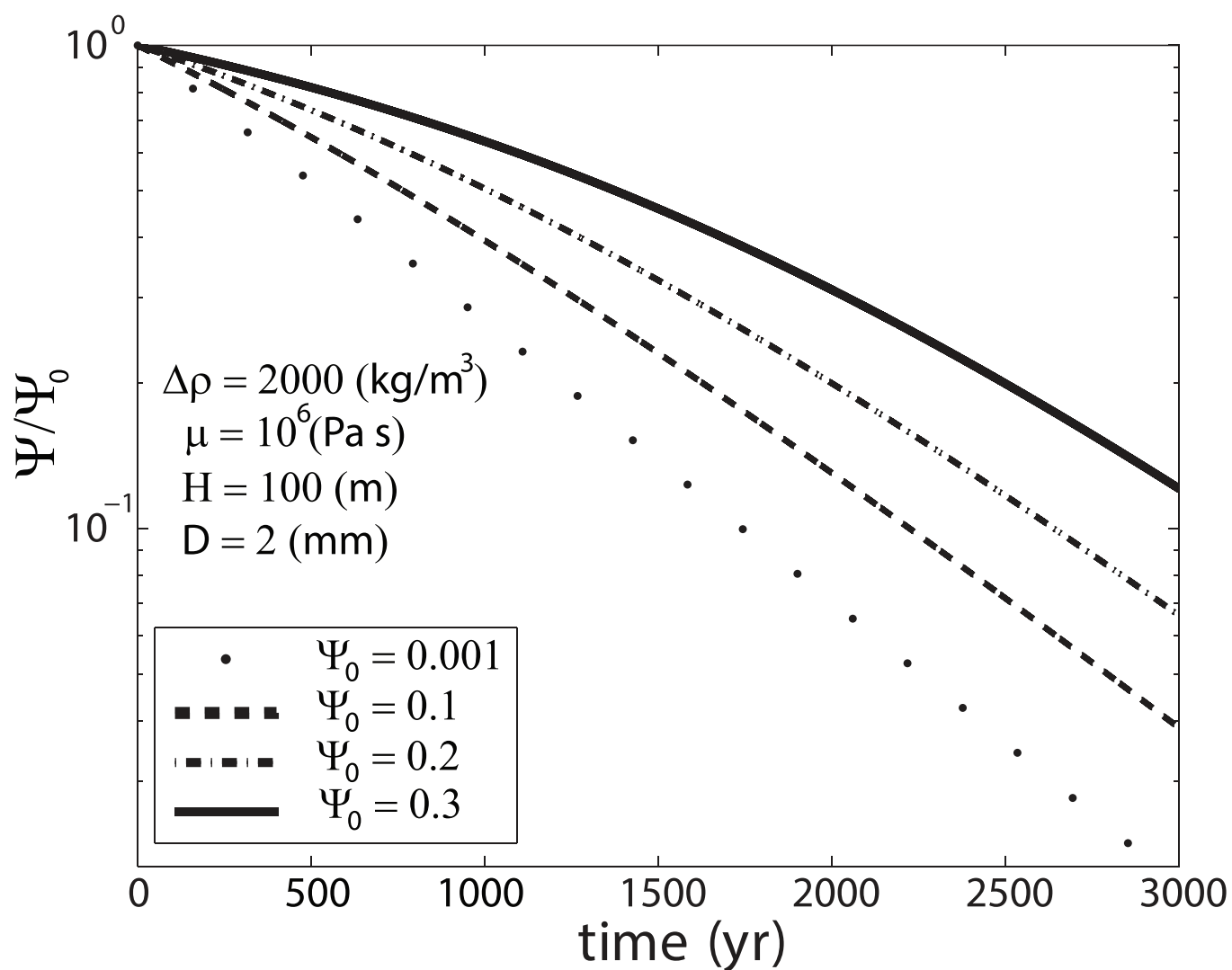


Extended Data Figure 2 | Confinement effect and MVP percolation. **a–f**, The results of three numerical calculations used to explain the effect of crystal confinement on fingering formation (see video in Supplementary Information). Porosity, ϕ , decreases from left to right. **a–c**, Three separate initial states, at different porosities; **d–f**, the corresponding steady states, at the corresponding porosities. At higher crystallinity ($1-\phi$), fingers can form and remain stable. **g**, Results of 78 calculations showing the correlation between the MVP volume fraction,

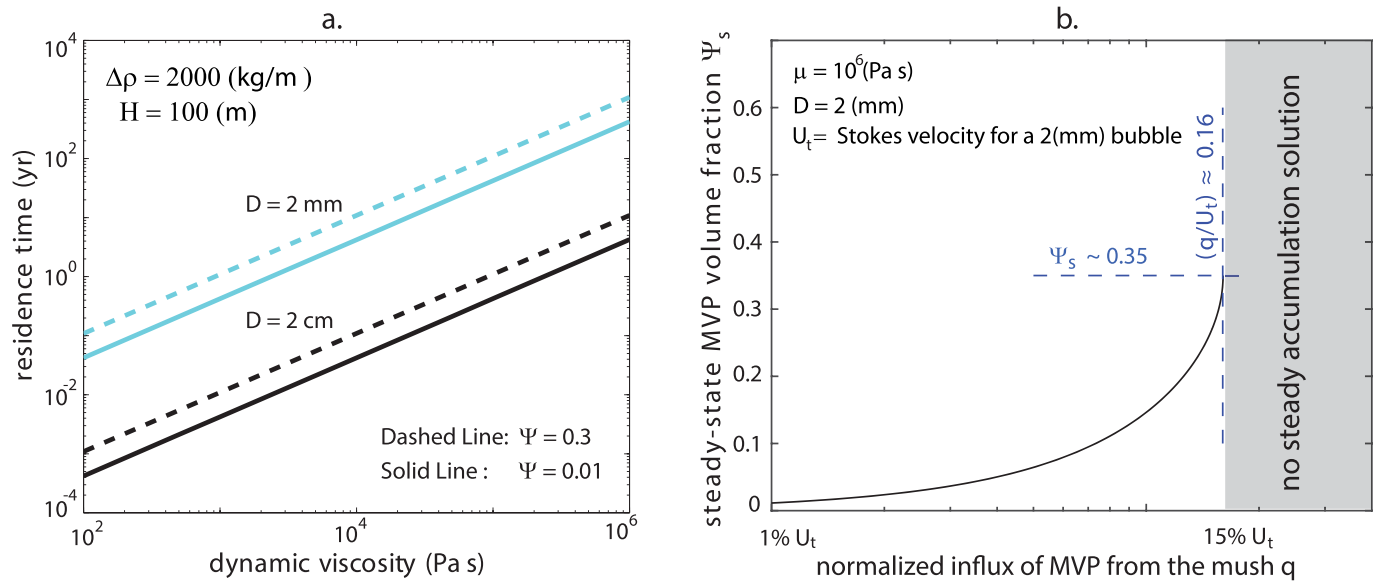
Ψ , and the flux of MVP in the porous medium (the Darcy velocity, U_{Darcy}). At low Ψ , the low mobility of bubbles is such that U_{Darcy} is close to zero. Once continuous fingers are formed ('connected'; green and blue regions), the MVP flux experiences a strong increase because of the sudden and sharp decrease in the rate of viscous energy dissipation. Conversely, during a waning influx of MVP (moving from right to left in **g**), an MVP volume fraction of 10% or slightly more can remain trapped in the mush because of capillary and viscous trapping in the mush.



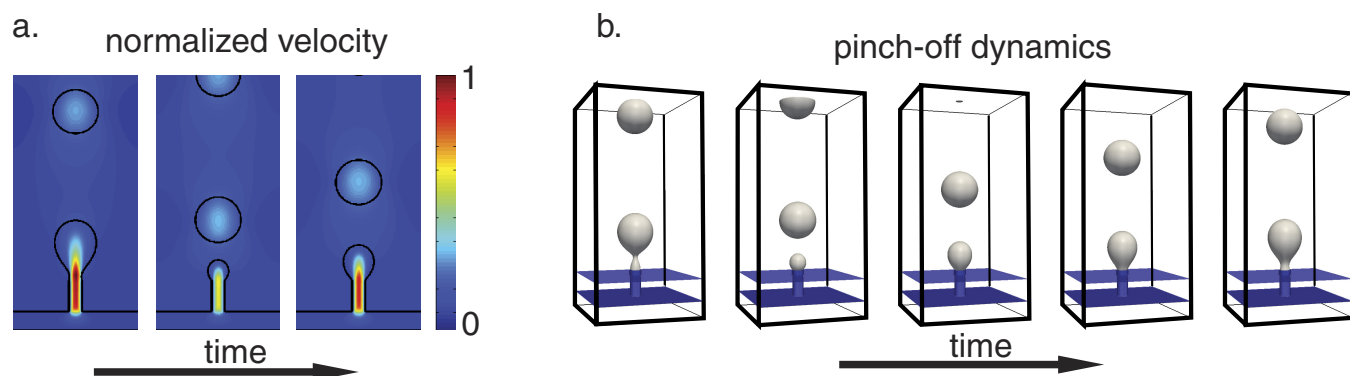
Extended Data Figure 3 | Experimental study of bubble separation in suspensions. Water droplets are released from localized nozzles at the top and sink into viscous silicon oil, forming bubble trains or plumes initially. The motion of water droplet is captured by a camera and used to test our bubble suspension migration model (equation (1)).



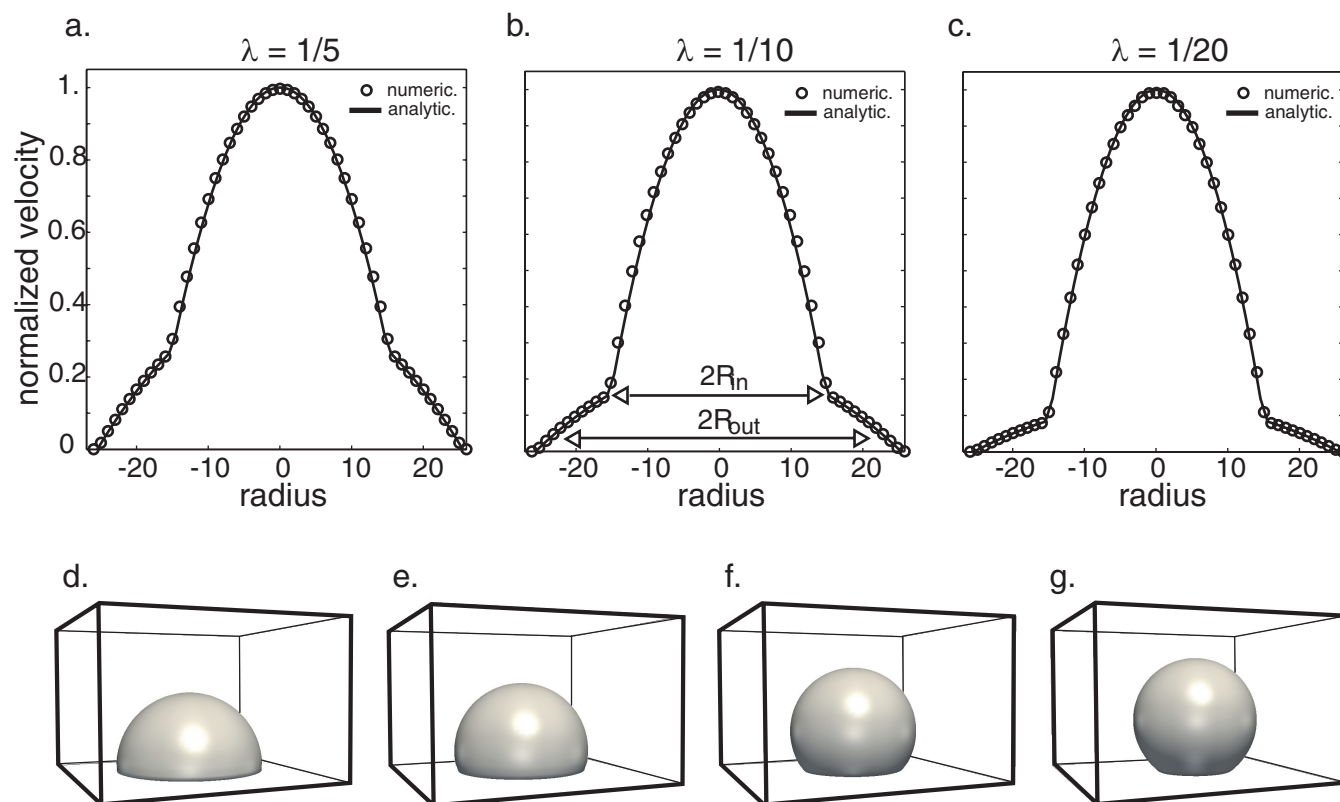
Extended Data Figure 4 | Residence time of bubbles in convecting crystal-poor magmas. For conditions and parameters consistent with exsolved volatile bubbles (2 mm diameter) in a viscous melt, the detrainment of bubbles over time depends on the initial bubble volume fraction, because of the hindered motion of bubbles in a suspension.



Extended Data Figure 5 | Bubble accumulation in convecting magma. **a, b,** Bubble residence time (**a**) and accumulation (**b**) in a convecting crystal-poor cap of thickness H (100 m). D refers to the average diameter of bubbles; $\Delta\rho$ is the density difference between MVP and the magma; q is the volumetric flux of MVP coming from the mush; and Ψ_s is the volume fraction of bubbles that can accumulate in the convecting layer.



Extended Data Figure 6 | Pinch-off dynamics. a, b, Results of numerical calculations that show the transition in transport regime of MVP from a confined medium (crystal-rich mush; left) to an unconfined horizon (crystal-poor cap; right).



Extended Data Figure 7 | Validation of the lattice Boltzmann algorithm: cylindrical Poiseuille flow and static contact angles. **a–c,** Analytical (equations (8) and (9)) and numerical velocity (lattice Boltzmann algorithm) profiles for a three-dimensional, two-immiscible-phase, cylindrical pipe flow scenario at different viscosity ratios ($\lambda=1/5$, $1/10$, or $1/20$), showing normalized bubble velocity versus pipe radius. A bulk force, F_b , is applied to both fluids. R_{in} and R_{out} are the internal and external

radius, respectively, for the annular flow. **d–g,** Different static contact angles obtained with our lattice Boltzmann algorithm. From left to right, we increase the non-wetting potential of the dispersed phase. The bubble contact angle accordingly increases from 90° to 150° (**d**, 90°; **e**, 110°; **f**, 130°; **g**, 150°). The calculations were done with an MRT collision operator (see Methods).

The ‘Tully monster’ is a vertebrate

Victoria E. McCoy¹, Erin E. Saupe¹, James C. Lamsdell^{1,2}, Lidya G. Tarhan¹, Sean McMahon¹, Scott Lidgard³, Paul Mayer³, Christopher D. Whalen¹, Carmen Soriano⁴, Lydia Finney⁴, Stefan Vogt⁴, Elizabeth G. Clark¹, Ross P. Anderson¹, Holger Petermann¹, Emma R. Locatelli¹ & Derek E. G. Briggs^{1,5}

Problematic fossils, extinct taxa of enigmatic morphology that cannot be assigned to a known major group, were once a major issue in palaeontology. A long-favoured solution to the ‘problem of the problematica’¹, particularly the ‘weird wonders’² of the Cambrian Burgess Shale, was to consider them representatives of extinct phyla. A combination of new evidence and modern approaches to phylogenetic analysis has now resolved the affinities of most of these forms. Perhaps the most notable exception is *Tullimonstrum gregarium*³, popularly known as the Tully monster, a large soft-bodied organism from the late Carboniferous Mazon Creek biota (approximately 309–307 million years ago) of Illinois, USA, which was designated the official state fossil of Illinois in 1989. Its phylogenetic position has remained uncertain and it has been compared with nemerteans^{4,5}, polychaetes⁴, gastropods⁴, conodonts⁶, and the stem arthropod *Opabinia*⁴. Here we review

the morphology of *Tullimonstrum* based on an analysis of more than 1,200 specimens. We find that the anterior proboscis ends in a buccal apparatus containing teeth, the eyes project laterally on a long rigid bar, and the elongate segmented body bears a caudal fin with dorsal and ventral lobes^{3–6}. We describe new evidence for a notochord, cartilaginous arcualia, gill pouches, articulations within the proboscis, and multiple tooth rows adjacent to the mouth. This combination of characters, supported by phylogenetic analysis, identifies *Tullimonstrum* as a vertebrate, and places it on the stem lineage to lampreys (Petromyzontida). In addition to increasing the known morphological disparity of extinct lampreys^{7–9}, a chordate affinity for *T. gregarium* resolves the nature of a soft-bodied fossil which has been debated for more than 50 years.

Since *T. gregarium* was originally described as a representative of an extinct phylum^{3,5}, there have been only two attempts using extensive

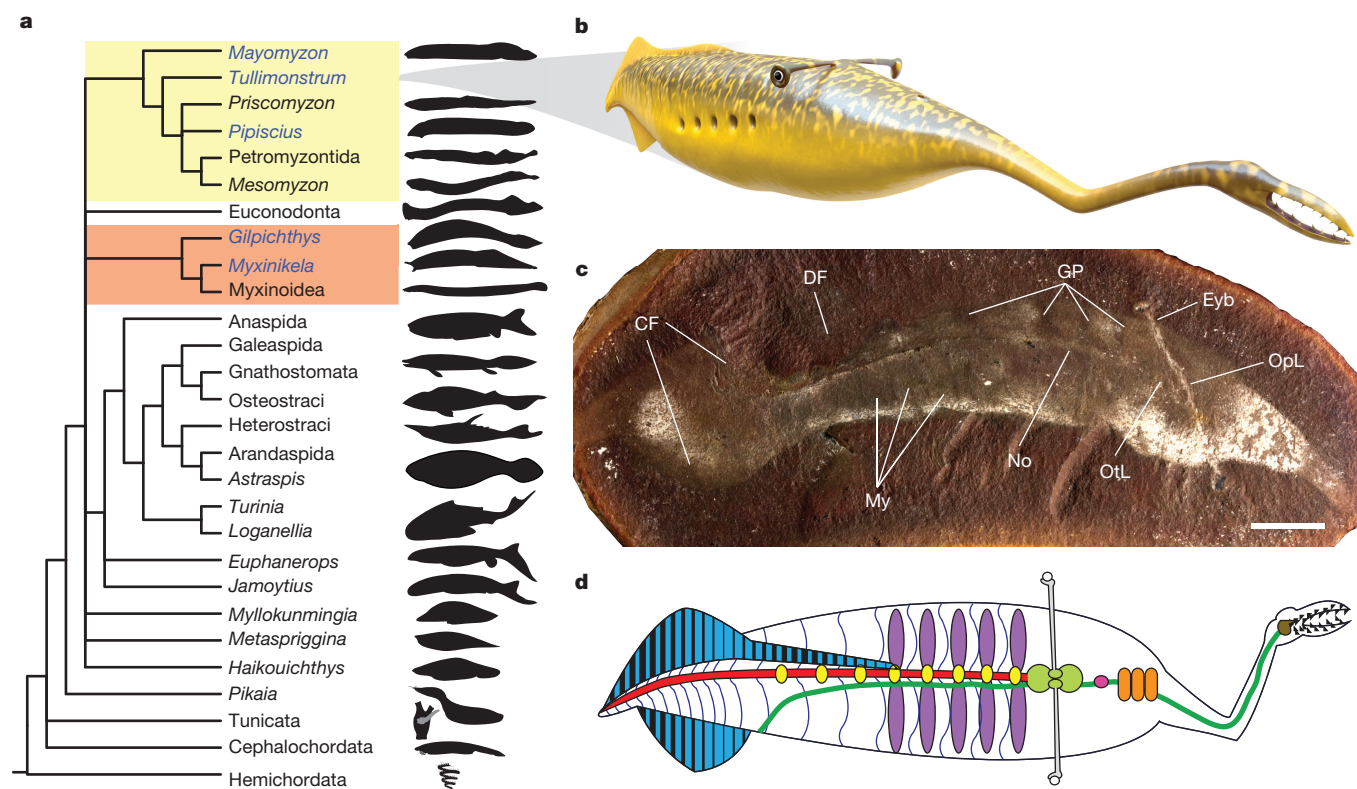


Figure 1 | Morphology and phylogeny of *Tullimonstrum*. **a**, Chordate phylogeny including *T. gregarium*; lampreys in yellow; hagfishes in orange. **b**, Reconstruction of *Tullimonstrum*. **c**, *Tullimonstrum*, FMNH PE 40113, oblique lateral view (also Extended Data Fig. 2a): eyebar, Eyb; myomeres, My; gill pouches, GP; caudal fin, CF; notochord, No; otic lobe, OtL and

optic lobe, OpL of brain; and dorsal fin, DF. **d**, Line drawing: black, teeth; brown, lingual organ; light grey, eyebar; dark green, gut and oesophagus; red, notochord; light green, brain; orange, tectal cartilages; pink, naris; purple, gill pouches; yellow, arcualia; dark blue, myosepta; blue with black stripes, fins with fin rays. Scale bar, 10 mm.

¹Department of Geology and Geophysics, Yale University, 210 Whitney Avenue, New Haven, Connecticut 06511, USA. ²American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024, USA. ³Field Museum of Natural History, 1400 S. Lake Shore Drive, Chicago, Illinois 60605, USA. ⁴X-ray Science Division, Advanced Photon Source, Argonne National Laboratory, Argonne, Illinois 60439, USA. ⁵Yale Peabody Museum of Natural History, 170 Whitney Avenue, New Haven, Connecticut 06511, USA.



Figure 2 | Notochord and gut trace of *T. gregarium* and *G. greenei*. a–d, White arrows, notochord; black arrows, gut trace. Scale bars, 10 mm, except in inset in c which is 1 mm. a, FMNH PE 22077, *Tullimonstrum*, dorsal view. b, FMNH PF8349, *Gilpichthys*, lateral view. c, FMNH PE9864, *Tullimonstrum*, oblique lateral view, with inset showing gut trace. See also Extended Data Fig. 2b. d, FMNH PF8480, *Gilpichthys*, lateral view.

additional specimens to resolve its affinity^{4,6}. These analyses favoured its interpretation as a swimming gastropod similar to living heteropods^{4,6}, or as a chordate close to conodonts⁶. Distinguishing between these alternatives depends primarily on the interpretation of three major morphological features, which are generally referred to as the (1) ‘gut trace’^{4,5}, a two-dimensional, light-coloured medial structure (Figs 1c and 2a), (2) ‘segments’^{5,6}, which consist of regularly spaced dark and light transverse bands (Figs 1c, and 2a and Extended Data Fig. 1a–g), and (3) ‘jaw apparatus’^{4–6} (Fig. 3a, b).

Our new investigation of >1,200 specimens (Extended Data Table 1) of *Tullimonstrum* counters the interpretation^{4,5} of the medial structure as a ‘gut trace’. It is preserved differently to the gut in other Mazon Creek animals, which is most commonly three-dimensional and filled by a dark mineral (Fig. 2d). The medial structure is preserved in a manner that most closely resembles the notochord of the stem hagfish *Gilpichthys greeniei*^{7,10,11} (Fig. 1a) from Mazon Creek, which is also a two-dimensional, light-coloured structure¹⁰ (Fig. 2b). *Gilpichthys* preserves a gut trace with the typical dark, three-dimensional

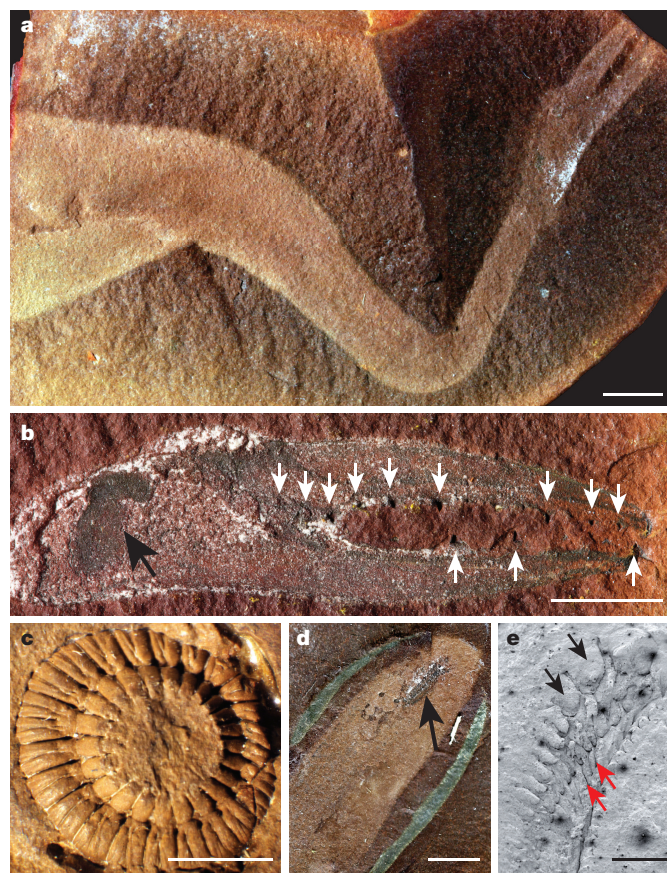


Figure 3 | Buccal apparatuses of jawless vertebrates at Mazon Creek.

a, b, *Tullimonstrum*. a, FMNH PE39375, jointed proboscis. Scale bar, 10 mm. b, FMNH PE45419, buccal apparatus (bifurcate structure and base), lingual organ (arrow), and pyritized teeth (white arrows) that continue into the base (see Extended Data Fig. 4a). Scale bar, 5 mm. c, *Pipiscius*, FMNH PF8344, concentric circles of plates. Scale bar, 1 mm. d, e, *Gilpichthys*. d, FMNH PF8480, buccal apparatus (arrow). Scale bar, 5 mm. e, FMNH PF8420, scanning electron microscope (SEM) image of buccal apparatus with muscle blocks (black arrows) and teeth (red arrows). Scale bar, 1 mm.

preservation¹⁰ (Fig. 2d), as well as the notochord. A true gut trace is evident in a few specimens of *Tullimonstrum* (Fig. 2c) where it ends in the expected position just anterior to the tail (Extended Data Fig. 2b). This feature is ventral in position in specimens of *Tullimonstrum* that afford a lateral view (Supplementary Information), in contrast to the light-coloured linear structure which is in the expected dorsal position of a notochord and continues, unlike the gut trace, into the tail (Extended Data Fig. 1f)⁹. The notochord of *Tullimonstrum* lies immediately posterior to a three-lobed structure (in some specimens the two features appear continuous) around the eyebar, which is very similar to the tri-lobed chordate brain⁹. The eyebar connects the eyes to the central (optic) lobe, suggesting that the eyebar protects the optic nerves (Extended Data Fig. 3a–d). A series of ‘medial organs’^{4,5} (Extended Data Fig. 1d, g) is associated with the notochord (Extended Data Fig. 2c). Their three-dimensional preservation suggests that they were relatively decay resistant^{5,12} and we interpret them as cartilaginous arcualia. Similar (but smaller and less regular) three-dimensional, repeated structures also occur along the notochord of *Gilpichthys*¹⁰. Serially arranged internal structures are also present in some molluscs but they are almost always paired¹³ and are therefore unlike the structures interpreted here as arcualia in *Tullimonstrum*.

The ‘segments’^{5,6} of *Tullimonstrum* (Extended Data Fig. 1a–g) are muscle blocks which have separated as a result of decay. They are most commonly preserved in the shape of a W or chevron, but they may also be straight (Extended Data Fig. 1a–g), and one arcualium corresponds

to each 'segment'⁵ (Extended Data Fig. 1d, g). The variable morphologies of these features fall within the range generated by the decay of chordate myomeres¹⁴. Newly observed structures in the anterior region of the body in some specimens of *Tullimonstrum*, spaced in a similar manner to the myomeres, appear to represent gill pouches (Fig. 1c and Extended Data Fig. 1h). These structures are typically wider than the myomere separations and elliptical in shape; up to five are evident (Extended Data Fig. 1h). The myomere separations are most pronounced at the edge of the body whereas the gill pouches lie close to the mid-line (Extended Data Fig. 1g, h).

The body of *Tullimonstrum* extends anteriorly into a proboscis terminating in the jaw apparatus^{4,5}, which includes an asymmetric proximal base and a distal bifurcate projection bearing small, pointed 'stylets'⁵ (Fig. 3a, b). Individual stylets show a range of morphologies which reflect variable cross-sections through a slightly hooked, hollow cone⁴ (Extended Data Figs 4d–i and 5m, n). Differences in preservation suggest that the jaw apparatus was more decay resistant than the body, and that the stylets were composed of the most recalcitrant material of the three^{4,5}. The jaw apparatus was previously interpreted as a bifurcate buccal apparatus bearing teeth⁵, or as the cross section of a hollow cylindrical buccal mass containing a radula^{4,6}. Our new observations resolve its nature and function.

The stylets, here interpreted as teeth, form at least two rows on dorsal and ventral parts of the bifurcate structure (Extended Data Fig. 4b). Each tooth is set in a small projection of soft tissue (Fig. 3b and Extended Data Fig. 4a). The teeth curve and project posteriorly, presumably to prevent food from escaping as it was ingested. The rows of teeth extend from the anterior tip of the bifurcate structure into the base (Fig. 3b and Extended Data Fig. 4a), indicating that the mouth opening was located there. The position of the gut is evident inside the proboscis⁵ (Extended Data Fig. 3d and Supplementary Information). The teeth in *Tullimonstrum* are very similar to those of lampreys and hagfish, which are often simple, posteriorly pointed, keratinous cones, set in raised soft tissue on a cartilage-supported base^{15,16}. The molluscan radula, in contrast, is a ribbon with complex, chitinous teeth^{4,17} which are rarely simple cones⁴ and are of the same composition as the tissue that bears them (Extended Data Fig. 6d).

The proboscis is not, as commonly described, flexible⁴, but is instead characterized by three distinct articulations (Fig. 3a and Supplementary Information): proximally at the connection to the head, about the midlength, and distally where it connects to the base of the buccal apparatus. Unlike Mazon Creek instances of hard mineralized tissues such as bone and shell¹², these articulations are not three-dimensionally preserved, and were probably supported by unmineralized cartilage.

An elliptical to circular dark stain situated peripherally within the asymmetric base of the buccal apparatus (Fig. 3b and Extended Data Fig. 4a, b) is interpreted as decay-resistant internal tissue, either cartilage or muscle, which bears some resemblance in morphology and position to the lamprey lingual organ¹⁵.

A phylogenetic analysis of early chordates¹⁸, together with *Tullimonstrum*, resolves it within the lamprey stem lineage (Fig. 1 and Extended Data Fig. 7). A number of features align *Tullimonstrum* with lampreys (Supplementary Information): pronounced cartilaginous arcualia; a dorsal fin and asymmetric caudal fin; keratinous teeth; a single nostril; and the presence of tectal cartilages (Extended Data Fig. 8b).

Unsurprisingly, given the difficulty of determining its affinities, *Tullimonstrum* displays several features that are not found in lampreys, but these probably reflect its distinct mode of life. There are 20–25 myomeres in the trunk and tail of *Tullimonstrum* (Extended Data Fig. 1a–c), compared with the 50–70 typical of lampreys¹⁹. A low myomere count is often associated with a short, stout body, a tail with a high caudal fin aspect ratio, and tail-propelled rather than undulatory swimming²⁰ (Extended Data Table 2 and Supplementary Information). Extant lampreys are undulatory swimmers, but the body shape of

Tullimonstrum suggests that it may have approached tail-propelled swimming.

The eyes of *Tullimonstrum* are set on a rigid horizontal bar (Fig. 1b–d), a configuration rare in chordates but present in hammerhead sharks and larval dragon fish. The eye position is too poorly resolved to reconstruct the visual field of *Tullimonstrum*, but may reflect the position of the prey-capturing buccal apparatus at the end of a long anterior extension of the head.

The buccal apparatus of *Tullimonstrum* suggests that it grasped food with its bifurcate anterior projection, and rasped pieces off with the lingual apparatus¹⁵ (Extended Data Fig. 4a, b). *Tullimonstrum* is one representative of a diverse fauna of Mazon Creek jawless vertebrates with a variety of feeding structures²¹: stem lampreys *Mayomyzon piekoensis*⁸ and *Pipiscius zangerli*^{7,10,11}, and stem hagfishes *Myxinikela siroka*²² and *G. greeni*^{7,10,11} (our phylogenetic analysis supports the assignment of *Pipiscius* and *Gilpichthys* to stem lampreys and stem hagfishes, respectively) (Fig. 1a). *Pipiscius* has a unique buccal apparatus consisting of two concentric circles of plates around an enlarged pharyngeal chamber¹⁰ (Fig. 3c). *Gilpichthys* also has a unique buccal apparatus, consisting of an elongate pharyngeal chamber lined with muscle blocks bearing posteriorly directed teeth¹⁰ (Fig. 3d, e). *Tullimonstrum* significantly expands the morphological disparity known in the lamprey lineage, providing insight into a clade that is characterized by highly conserved morphologies today.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 14 October 2015; accepted 12 January 2016.

Published online 16 March 2016.

1. Bengtson, S. in *Problematic Fossil Taxa* (eds Hoffman, A. & Nitecki, M. H.) 3–11 (Oxford Univ. Press, 1986).
2. Gould, S. J. *Wonderful Life: The Burgess Shale and the Nature of History* (WW Norton, 1990).
3. Richardson, E. S. Jr. Wormlike fossil from the Pennsylvanian of Illinois. *Science* **151**, 75–76 (1966).
4. Foster, M. in *Mazon Creek Fossils* (ed. Nitecki, M. H.) 269–301 (Academic, 1979).
5. Johnson, R. G. & Richardson, E. S. Pennsylvanian invertebrates of the Mazon Creek Area, Illinois: the morphology and affinities of *Tullimonstrum*. *Fieldiana Geol.* **12**, 119–149 (1969).
6. Beall, B. in *The Early Evolution of Metazoa and the Significance of Problematic Taxa* (eds Simonetta, A. M. & Conway Morris, S.) 271–286 (Cambridge Univ. Press, 1991).
7. Janvier, P. & Sansom, R. S. in *Hagfish Biology* (eds Edwards, S. L. & Goss, G. G.) 73–94 (CRC, 2015).
8. Bardack, D. & Zangerl, R. in *The Biology of Lampreys* (eds Hardisty, M. W. & Potter, I. C.) Vol. 1, 67–84 (Academic, 1971).
9. Janvier, P. *Early Vertebrates* (Oxford Monographs on Geology and Geophysics Vol. 33) (Oxford Univ. Press, 1996).
10. Bardack, D. & Richardson, E. Jr. New agnathous fishes from the Pennsylvanian of Illinois. *Fieldiana Geol.* **33**, 489–510 (1977).
11. Janvier, P. The phylogeny of the Craniata, with particular reference to the significance of fossil "agnathans". *J. Vert. Paleont.* **1**, 121–159 (1981).
12. Baird, G. C., Siroka, S. D., Shabica, C. W. & Kuecher, G. J. Taphonomy of Middle Pennsylvanian Mazon Creek area fossil localities, northeast Illinois: significance of exceptional fossil preservation in syngenetic concretions. *Palaio* **1**, 271–285 (1986).
13. Jacobs, D. K. et al. Molluscan engrailed expression, serial organization, and shell evolution. *Evol. Dev.* **2**, 340–347 (2000).
14. Sansom, R. S., Gabbott, S. E. & Purnell, M. A. Atlas of vertebrate decay: a visual and taphonomic guide to fossil interpretation. *Palaentology* **56**, 457–474 (2013).
15. Yalden, D. Feeding mechanisms as evidence for cyclostome monophyly. *Zool. J. Linn. Soc.* **84**, 291–300 (1985).
16. Alibardi, L. & Segalla, A. The process of cornification in the horny teeth of the lamprey involves proteins in the keratin range and other keratin-associated proteins. *Zool. Stud.* **50**, 416–425 (2011).
17. Scheltema, A. H., Kerth, K. & Kuzirian, A. M. Original molluscan radula: comparisons among Aplousobranchia, Polyplacophora, Gastropoda, and the Cambrian fossil *Wiwaxia corrugata*. *J. Morphol.* **257**, 219–245 (2003).
18. Morris, S. C. & Caron, J.-B. A primitive fish from the Cambrian of North America. *Nature* **512**, 419–422 (2014).
19. Meeuwij, M. H., Bayer, J. M. & Reiche, R. A. Morphometric discrimination of early life stage *Lampetra tridentata* and *L. richardsoni* (Petromyzonidae) from the Columbia River Basin. *J. Morphol.* **267**, 623–633 (2006).

20. McDowall, R. M. Jordan's and other ecogeographical rules, and the vertebral number in fishes. *J. Biogeogr.* **35**, 501–508 (2008).
21. Janvier, P. Facts and fancies about early fossil chordates and vertebrates. *Nature* **520**, 483–489 (2015).
22. Bardack, D. First fossil hagfish (Myxinoidea): a record from the Pennsylvanian of Illinois. *Science* **254**, 701–703 (1991).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank affiliates of the Field Museum of Natural History: J. Voight and J. Wittry for discussion; P. Heck for help with scanning electron microscopy and energy-dispersive spectroscopy; and N. Karpus for specimen photography. This research used resources of the Advanced Photon Source, a US Department of Energy Office of Science User Facility operated for the US Department of Energy Office of Science by Argonne National Laboratory under contract number DE-AC02-06CH11357. The Field Museum of Natural History, the Sedgwick Museum in Cambridge, UK, and C. Eaton at the University of Wisconsin-Madison Geology Museum provided access to specimens. Access to the software TNT for phylogenetic analysis was provided

by the Willi Hennig Society. Funding was provided by a Field Museum visiting scholarship to V.E.M. and by the NASA Astrobiology Institute (NNA13AA90A) Foundations of Complex Life, Evolution, Preservation and Detection on Earth and Beyond.

Author Contributions V.E.M. conceived the study and wrote the initial draft. V.E.M., E.E.S., L.G.T., J.C.L., and D.E.G.B. developed the project. V.E.M., E.E.S., J.C.L., L.G.T., S.M., S.L., P.M., C.D.W., E.G.C., and R.P.A. analysed and measured specimens. J.C.L. ran the phylogenetic analysis. E.E.S., C.S., L.F., and S.V. performed the synchrotron analysis. S.M. created the reconstruction. H.P. dissected modern taxa for comparative purposes. E.R.L., E.E.S., and S.M. photographed comparative fossil taxa. All authors reviewed and edited the final manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to V.E.M. (victoria.mccoy@yale.edu).

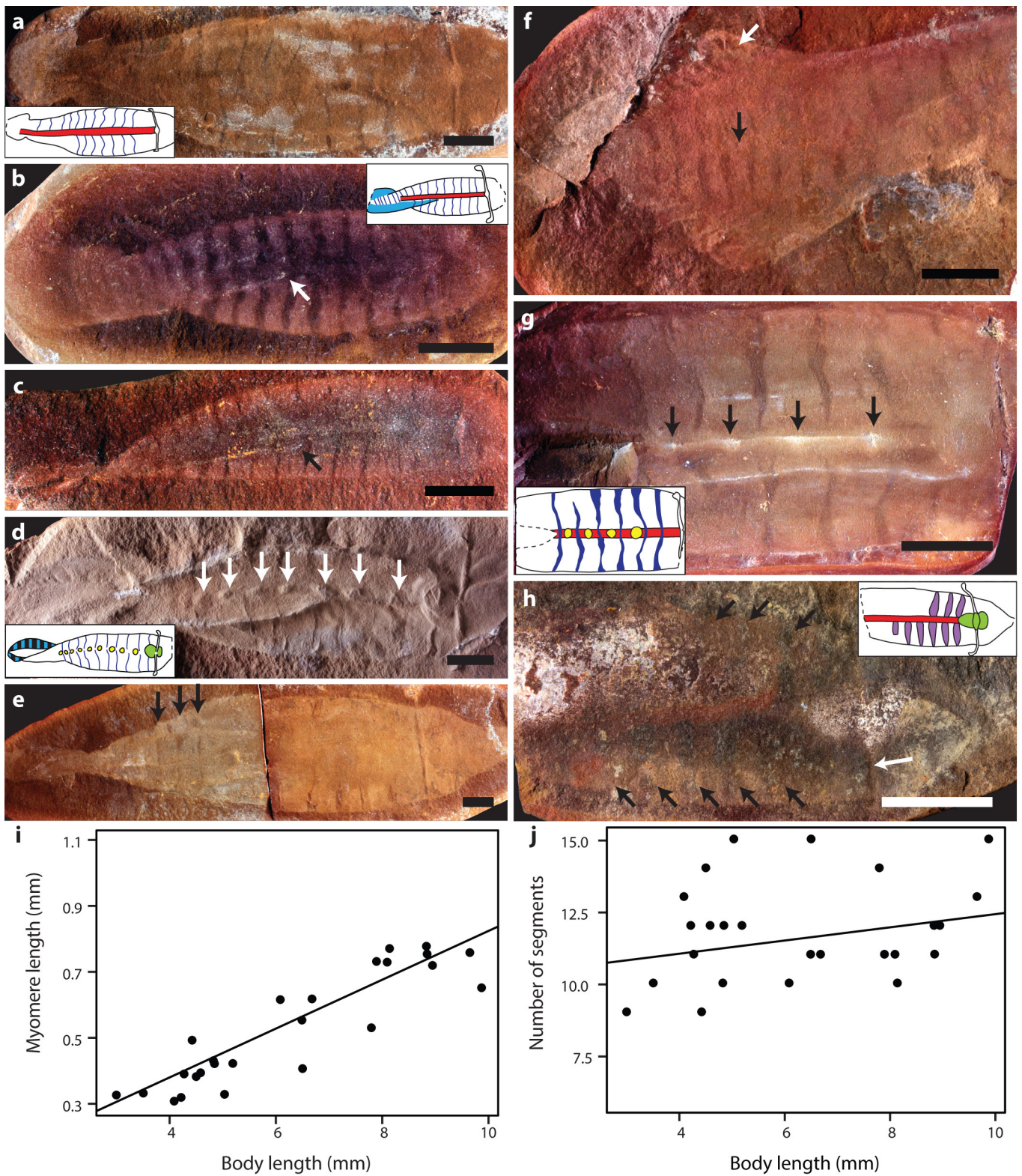
METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

All specimens figured and discussed here are held by the Field Museum of Natural History, Chicago, Illinois, USA. They were photographed in direct and polarized light, using a Canon EOS 60D camera. For some images, particularly of the buccal apparatus, multiple images were z-stacked in Helicon Focus version 4.2.8. All measurements were taken in ImageJ. Comparisons were made with specimens from the Sedgwick Museum, Cambridge, UK, and the University of Wisconsin-Madison Geology Museum. Specimens of *Tullimonstrum* and other taxa were analysed using energy-dispersive spectroscopy (EDS) on the SEM at the Field Museum of Natural History. Synchrotron images were collected at beamline 8-BM-B of the Advanced Photon Source. Incident X-rays (10.7 keV) were focused to a 30 µm beam spot using Kirkpatrick–Baez mirrors. The samples were raster-scanned through this spot, in steps of 0.02 mm (fine scans) to 0.2 mm (coarse scans). Full X-ray fluorescence emission spectra were collected at each position using a SII Vortex ME4 4-element silicon drift detector fitted with an aluminium filter to attenuate the strong iron signal of the matrix material, at 150 ms dwell per pixel. The spectra were fitted with Gaussian models of the characteristic energy and relative intensity of the known atomic emission peaks. The data were also normalized relative to incident flux, and counts were converted to a material quantity (micrograms per square centimetre) using relative calibration to thin-film AXO standards (AXO DRESDEN), with MAPS software²³. A modern lamprey (*Petromyzon marinus*), lancelet (*Branchiostoma* sp.), and hagfish (*Myxine glutinosa*) were dissected for anatomical comparison, with an emphasis on sectioning planes

that reflected the preservation of *Tullimonstrum*. *T. gregarium*, *G. greeni*, and *P. zangerli* were added to an existing phylogenetic matrix of basal vertebrates¹⁸. One new character, the presence of tectal cartilage, was added to the matrix. The only other changes to the matrix comprised alterations to the coding of Myxinoidea; specifically, characters 28 and 29 ('Single confluent branchial opening' and 'Elongate branchial series') were coded as polymorphic to represent the disparity in the clade of extant hagfishes. The resulting matrix of 117 characters and 28 taxa was analysed in TNT²⁴ using implicit enumeration with all characters unordered and of equal weight. Bootstrap²⁵, Jackknife²⁶, and Bremer²⁷ support values were calculated in TNT; the ensemble consistency, retention and rescaled consistency indices were calculated in Mesquite 3.02 (ref. 28). Bootstrapping was performed with 50% character resampling for 5,000 repetitions, and jackknifing by using simple addition sequence and tree bisection–reconnection branch swapping for 5,000 repetitions with 33% character deletion. The reconstruction of *Tullimonstrum* and its teeth was created with the free program Blender.

23. Vogt, S. MAPS: a set of software tools for analysis and visualization of 3D X-ray fluorescence data sets. *J. Phys. IV Fr.* **104**, 635–638 (2003).
24. Goloboff, P. A., Farris, J. A. & Nixon, K. C. TNT, a free program for phylogenetic analysis. *Cladistics* **24**, 774–786 (2008).
25. Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791 (1985).
26. Farris, J. S., Albert, V. A., Källersjö, M., Lipscomb, D. & Kluge, A. G. Parsimony jackknifing outperforms neighbor-joining. *Cladistics* **12**, 99–124 (1996).
27. Bremer, K. Branch support and tree stability. *Cladistics* **10**, 295–304 (1994).
28. Maddison, W. P. & Maddison, D. R. Mesquite: a modular system for evolutionary analysis v.3.02 (2015).

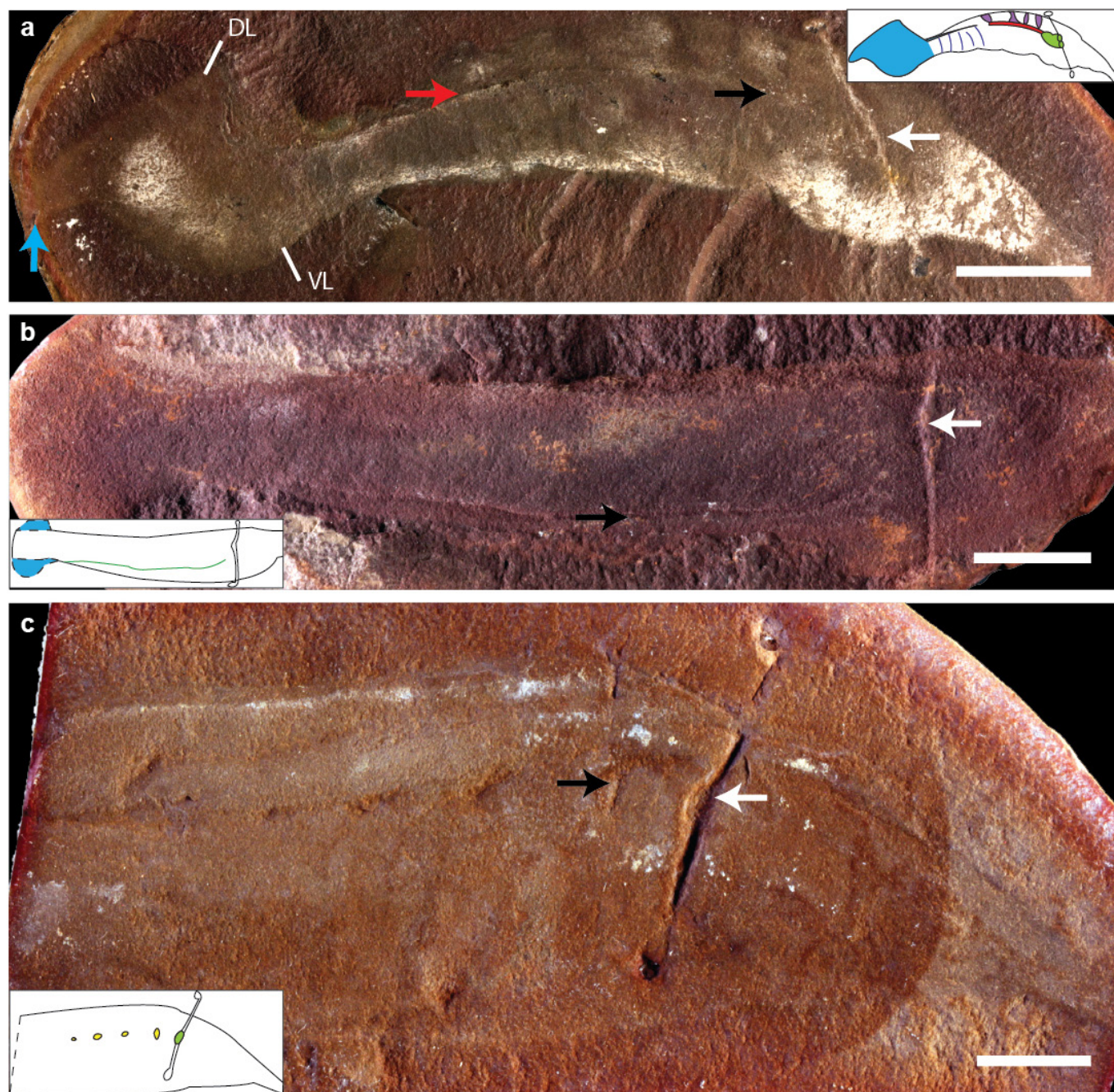


Extended Data Figure 1 | See next page for caption.

Extended Data Figure 1 | Details of myomeres, myomere separations, and gills in *Tullimonstrum*. Anterior to the right. Scale bars, 10 mm.

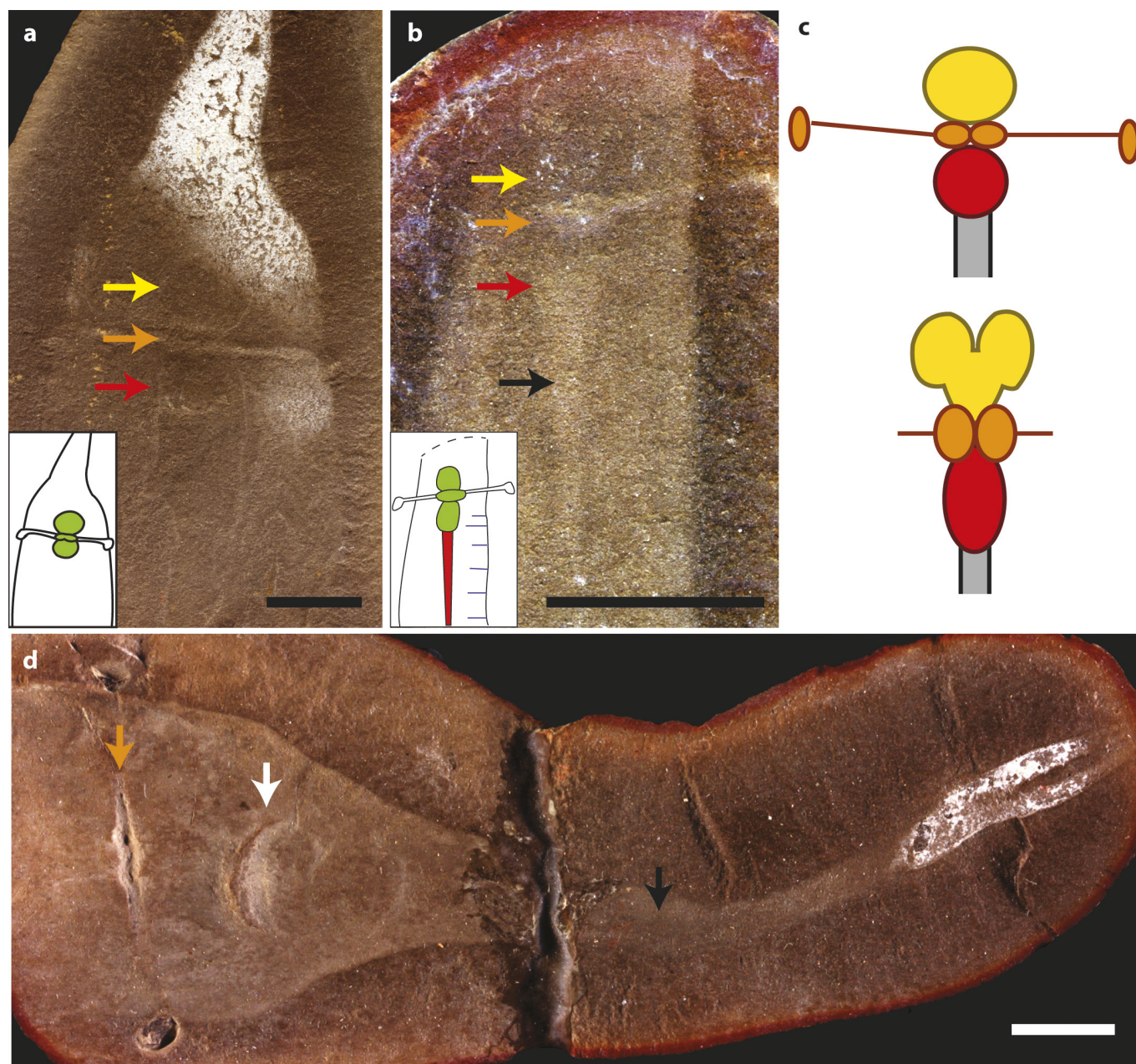
Insets are line drawings of the specimens, with colours as in Fig. 1d. **a**, **b**, W-shaped myomeres. **a**, FMNH PE32436, narrow myomere separations. **b**, FMNH PE32386, wider myomere separations. Note the folded tail and the dorsal fin which extends from approximately the fifth myomere (arrow). **c**, FMNH PE32423, chevron-shaped myomeres and dorsal fin extending from the fifth most posterior myomere (arrow). The tail appears unusually narrow because of an atypical vertical orientation. **d**, FMNH PE7063, straight myomeres separated only at the edges and corresponding to repeated arcualia (arrows). **e**, FMNH PE32395, myomeres separated only at the margins of the body except in the tail where they are completely separated (arrows). **f**, FMNH PE10654, myomeres separated in the tail. Note the axially positioned notochord in the tail (black arrow), and the asymmetric fin with rays (white arrow). **g**, FMNH PE10601, myomere

separations that resemble gill pouches but differ in their W-shape, lighter colour, and maximum width at the margin of the body. Note the arcualia (arrows) accentuated in white by kaolinite (arrows). **h**, FMNH PE45366, gill pouches (arrows) that are elliptical, darker in colour, and widest immediately adjacent to the notochord (reddish brown in the fossil). The eyebar is indicated by a white arrow. **i**, **j**, Linear regressions on segmentation variables. The assumptions for linear regression were tested using the R package *gvlma*, and all were met. The data are presented in Supplementary Table 2. Data were included for all specimens for which segments could be counted and measured, except for one outlier, which was removed from the calculations with no change to the *P* values or *R*² values of the regressions. **i**, Regression between body size and average myomere size, $n = 25$, $R^2 = 0.57$, $P = 4.92 \times 10^{-10}$. **j**, Regression between body size and number of myomeres in the body, $n = 25$, $R^2 = 0.05$, $P = 0.19$.



Extended Data Figure 2 | Dorsoventral position of axial structures in *Tullimonstrum*. Specimens preserved obliquely (as indicated by asymmetric preservation of the eyebar, its midpoint indicated by white arrow); the offset of medial structures relative to the axis of the specimen indicates their dorsoventral position. Anterior to the right. Insets are line drawings of the specimens, with colours as in Fig. 1d. Scale bars, 1 cm. **a**, FMNH PE40113, displacement of the centre of the eyebar, the notochord (black arrow), and the dorsal fin (red arrow) in the same direction, indicating that these are all dorsal structures. The notochord follows the curvature of the body. The tail bends ventrally at the posterior

tip (blue arrow) and the dorsal lobe (DL) of the caudal fin is longer than but not as deep as the ventral lobe (VL). **b**, **c**, Comparison of the position of additional medial features with that of the centre of the eyebar (which indicates the notochord position even when the notochord is not preserved). **b**, FMNH PE9864, displacement of the central bulb of the eyebar and the gut trace (black arrow), in opposite directions, indicating that the gut trace is ventral. **c**, FMNH PE24567, displacement of the central bulb of the eyebar and the arcualia (black arrow) in the same direction, indicating that the arcualia are dorsal.

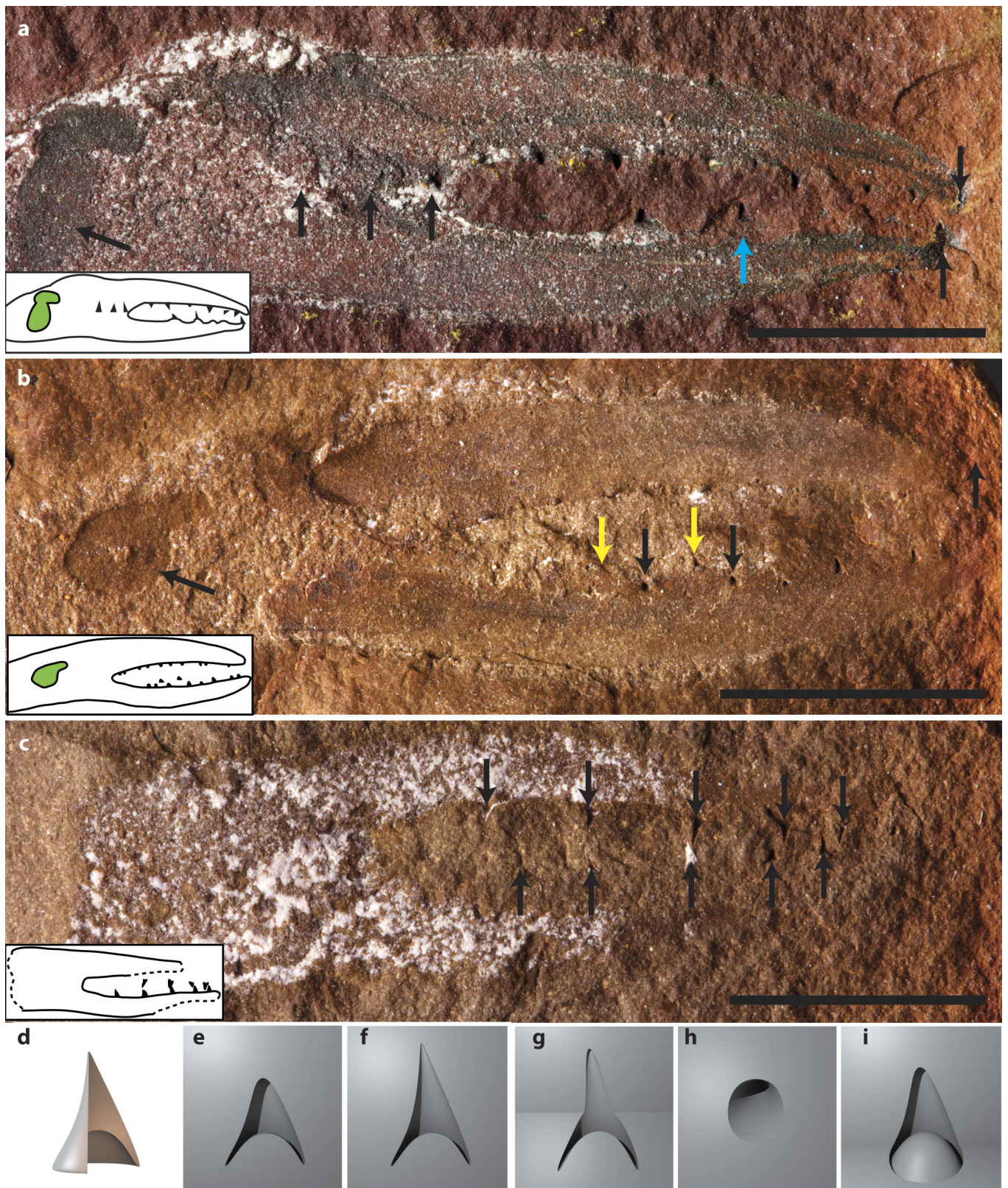


Extended Data Figure 3 | The tri-lobed brain of *Tullimonstrum*.

a–c, Anterior to the top; **d**, anterior to the right. Scale bars, 10 mm.

a, b, Yellow arrow, olfactory lobe; orange arrow, optic lobe; red arrow, otic lobe. **a**, FMNH PE45350, the three lobes of the brain, the large olfactory lobe anterior to the eyebar, the central optic lobe on the eyebar, and the large otic lobe posterior to it. **b**, FMNH PE22103, a faint trace of the anterior olfactory lobe, the central optic lobe preserved in association with the eyebar, and the posterior otic lobe immediately adjacent to the

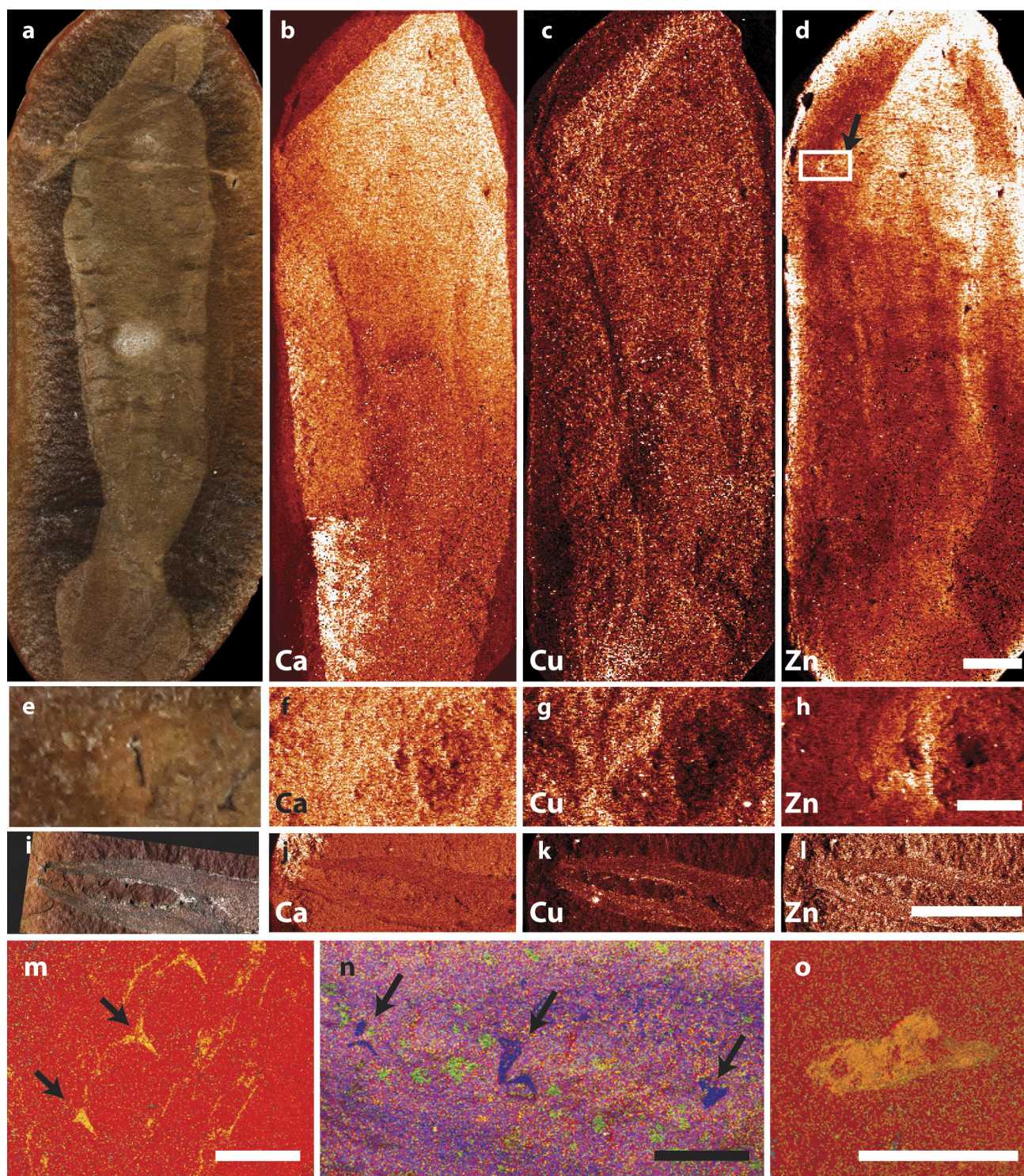
notochord (black arrow). **c**, Comparison of the brain of *Tullimonstrum* (above, based on FMNH PE45350 in **a**) with the brain of a typical lamprey (below) showing the olfactory lobe(s) in yellow, the optic lobes and optic nerve connections in orange, the cerebellum/medulla in red, and the notochord in grey. **d**, FMNH PE39890, bilobed central optic lobe (orange arrow) on the eyebar. Note the faint dark trace of the oesophagus in the proboscis (black arrow) and the crescent-shaped naris (white arrow).



Extended Data Figure 4 | See next page for caption.

Extended Data Figure 4 | Buccal apparatus of *Tullimonstrum*. All scale bars, 5 mm. Angled arrows indicate the lingual organ, and vertical arrows indicate teeth (not all of which are marked). **a**, FMNH PE45419, dark stain in the base of the apparatus that may represent a remnant of the lingual organ. The base and bifurcate structure are both asymmetric; the thicker (presumably dorsal) element of the bifurcate structure occurs on the same side as the dorsal bulge in the base. Three teeth (indicated by arrows) lie within the base rather than along the bifurcate structure. Teeth are present as far as the distal-most end of the bifurcate structure. The teeth are preserved as three-dimensional moulds and casts, occasionally as pyrite infills, and are situated on raised soft tissue areas (blue arrow in bifurcate structure). **b**, FMNH PE28739, lingual organ and two rows of teeth associated with the ventral element of the bifurcate structure. Two teeth in each row are indicated by arrows; the rows are offset and indicated

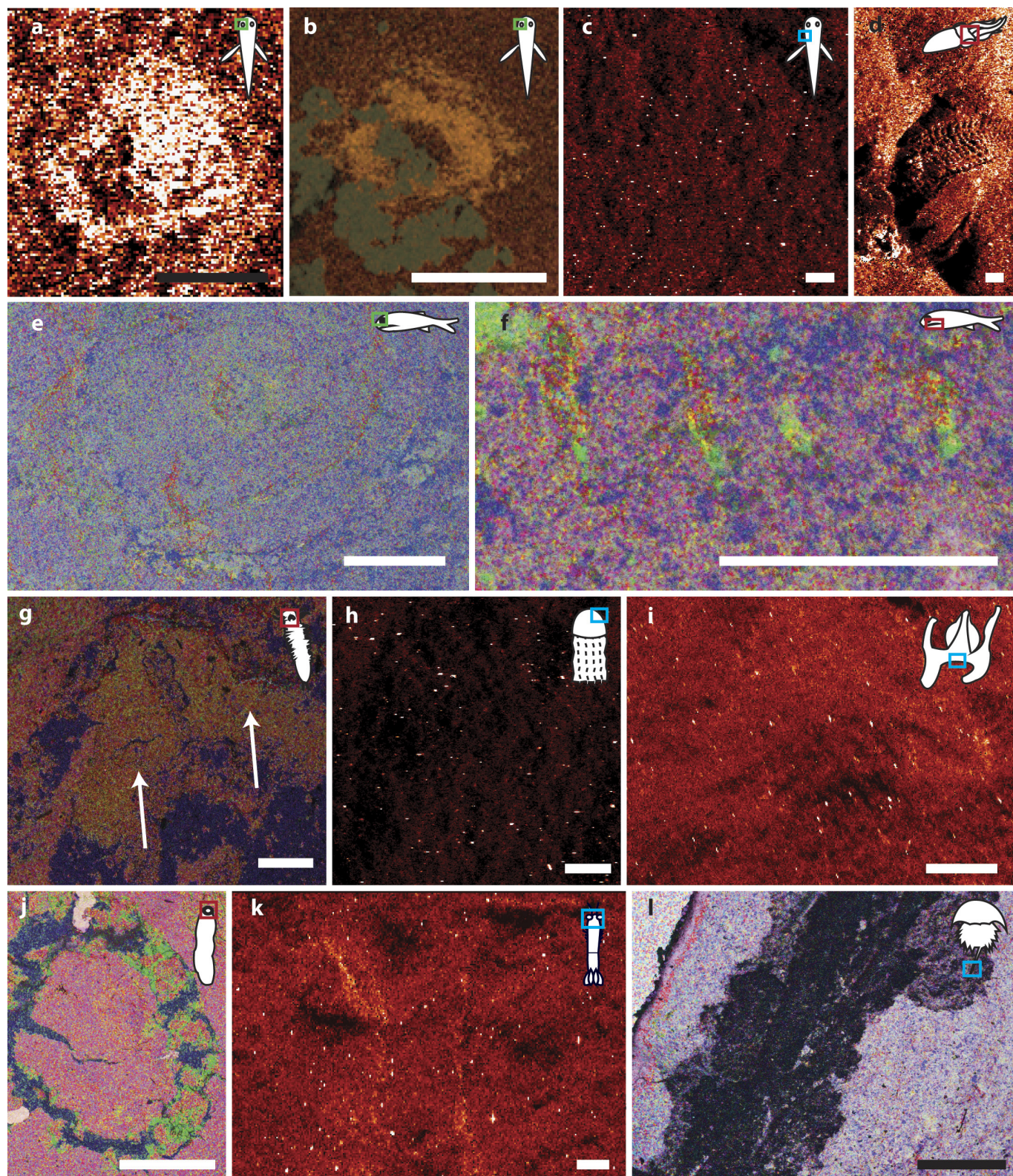
by black and yellow arrows. Teeth are preserved as moulds. **c**, FMNH PE31057, teeth with white kaolinite infill of negative-relief moulds. Other teeth are preserved as casts or moulds, beyond the distal-most preserved outline of the bifurcate structure. At the distal-most end, the teeth of each element of the bifurcate structure are offset (alternating), whereas for most of the length of the buccal apparatus they are not offset. The apparent asymmetry of the elements of the buccal apparatus is exaggerated as the kaolinite does not reveal the complete outline. **d**, A *Tullimonstrum* tooth in three dimensions as a slightly hooked hollow cone with a bulbous base. **e–i**, The three-dimensional tooth in **d** is digitally ‘rotated’ and ‘sliced’, to show how a two-dimensional representation of a three-dimensional structure may result in a variety of morphologies. These morphologies match those seen in preserved *Tullimonstrum* teeth, suggesting that the three-dimensional reconstruction in **d** might represent the original shape.



Extended Data Figure 5 | Elemental maps of *Tullimonstrum*.

a, e, i, Photographs and **(b–d, f–h)** synchrotron analysis of FMNH PE10504 and **(j–l)** FMNH PE45419, showing the distribution of calcium, copper, and zinc in **(b–d)** the body, **(f–h)** the eye (corresponding to the box **d**), and **(j–l)** the buccal mass. Scale bars, 10 mm (**a–d, i–l**) and 1 mm (**e–h, m–o**). Zinc shows a high concentration in the body and eye, but the buccal apparatus is characterized by enrichment in copper relative to the matrix. Note that the imaged buccal apparatus bears teeth, but they

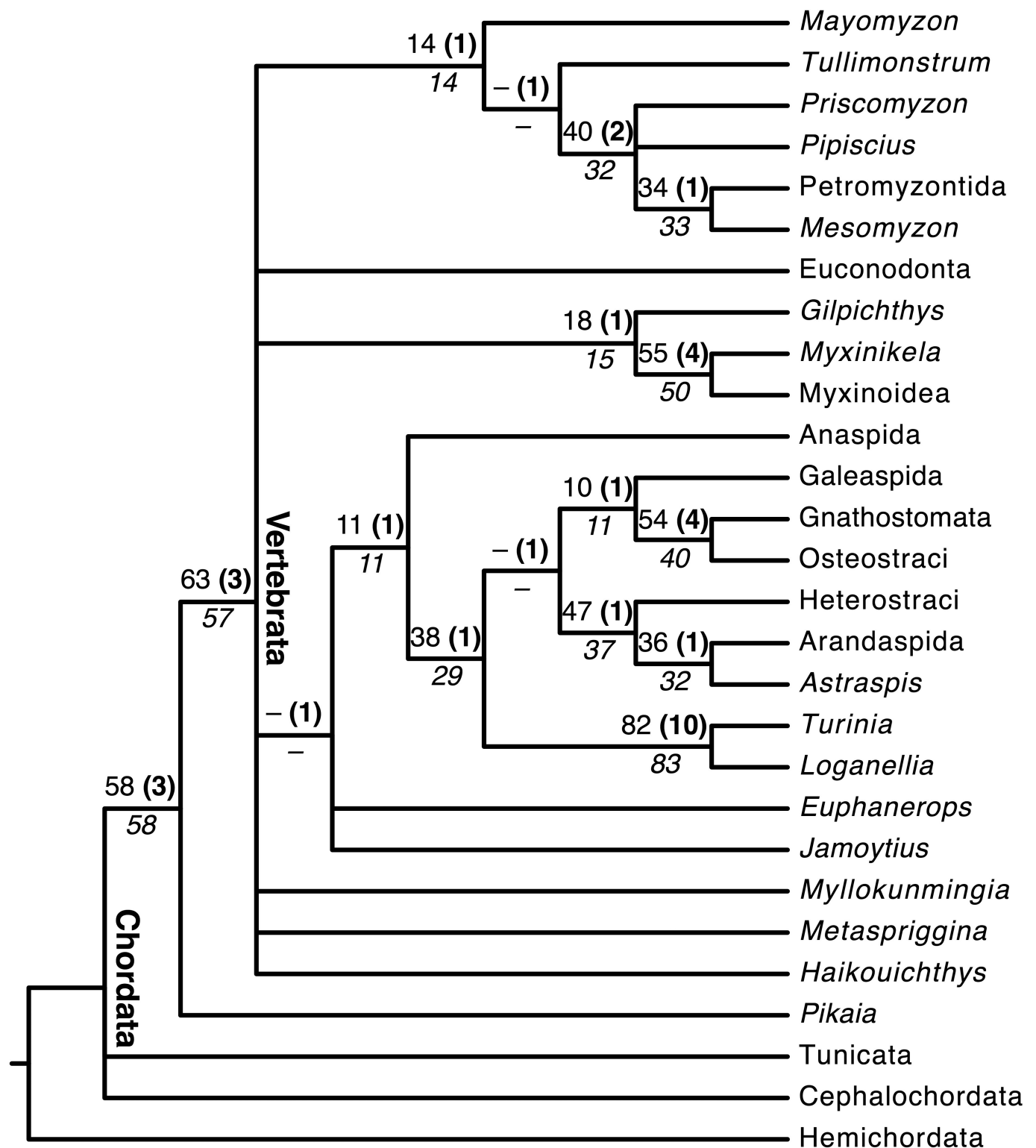
are characterized by enrichments distinct from those of the bifurcate structure. **m–o**, SEM/EDS analysis of **(m)** the teeth (at the spot indicated by the black arrow in **d**, where the proboscis folds back over the body) and **(o)** the eye of FMNH PE10504: yellow, sulphur; red, iron; showing the pyrite preservation (teeth indicated by arrows), and **(n)** the teeth of FMNH PE45426: green, calcium; blue, aluminium; pink, silicon; indicating clay mineral preservation of the teeth (in purple, indicated by arrows).



Extended Data Figure 6 | See next page for caption.

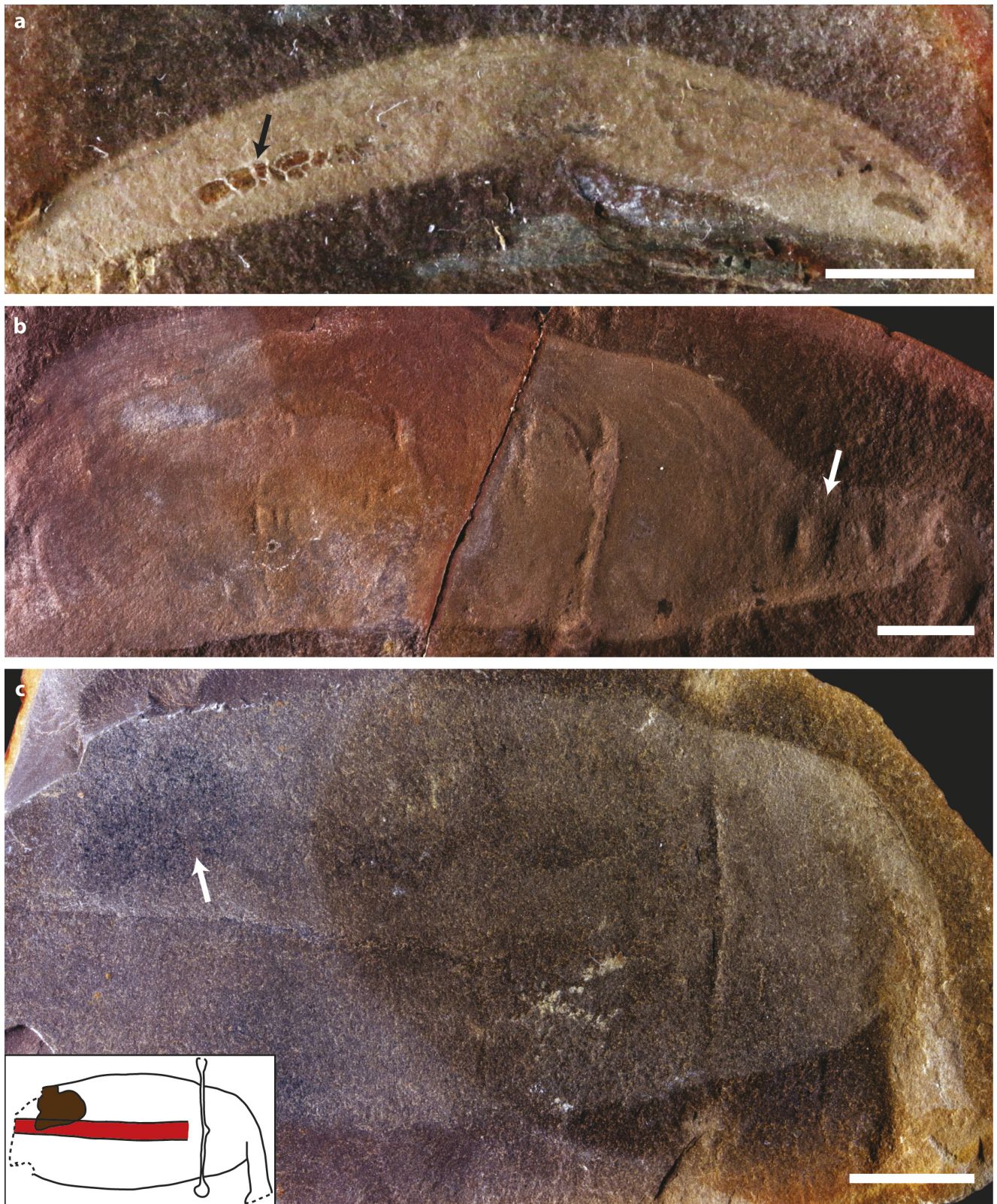
Extended Data Figure 6 | Elemental maps of comparative taxa. For each taxon, an icon shows the specimen regions analysed for Zn enrichment (brighter regions) with synchrotron (**a, c, d, h, i, k**) or for general elemental composition with SEM/EDS (**b, e–g, j, l**): green box, eyes; blue box, soft body tissue; red box, feeding structures. In SEM/EDS maps (**b, e–g, j, l**): red, iron; yellow, sulphur; green, calcium; pink, silicon; purple, phosphorus; blue, aluminium. The EDS cannot capture carbon, so a thin layer of it appears identical to the matrix, and a thick layer appears black. All scale bars 1 mm, except **f** which is 0.5 mm. **a–c**, FMNH PF8719, chordate *Esconichthys apopyris*; (**a**) eyes enriched in zinc, (**b**) preserved in pyrite, and (**c**) body lacking zinc enrichment. **d**, FMNH PE 39321, unidentified cephalopod, radula and soft tissue lacking zinc enrichment (dark areas are due to relief). **e, f**, FMNH PF7499, chordate *Elonichthys*

peltigerus; (**e**) eyes lacking pyrite and (**f**) teeth preserved in calcium, sulfur, and phosphorus interpreted to reflect the original apatite. **g**, FMNH PE12371, polychaete *Fossundecima konecniorum*, jaw apparatus (two light-coloured triangular areas, each with an arrow, on a darker background) preserved in a thin carbon film. **h**, FMNH PE30643, cnidarian *Essexella asherae*, tissue lacking zinc enrichment. **i**, FMNH PE 30054, *Etacystis communis*, unidentified soft-bodied organism, possibly cnidarian, soft tissue may be slightly enriched in zinc. **j**, FMNH PE21951, holothurian *Achistrum* sp., teeth of circular jaw preserved in calcium reflecting the original (calcite) composition. **k**, FMNH PE15530, crustacean *Lobetelson mclaughlinae*, eyes lacking zinc enrichment. **l**, FMNH PE23336, horseshoe crab *Euproops danae*, telson preserved as a thick carbon film.



Extended Data Figure 7 | Phylogenetic analysis. Strict consensus of the 18 most parsimonious trees retrieved under equal weighting (consistency index = 0.549, retention index = 0.614, rescaled consistency

index = 0.337). Jackknife support above each node in plain text, Bremer support emboldened and in parentheses, and bootstrap values beneath each node in italics.



Extended Data Figure 8 | Internal structures of *Gilpichthys* and *Tullimonstrum*. Anterior to right. **a**, FMNH PF8410, *Gilpichthys*, gut (arrow) preserved in a fashion similar to typical Mazon Creek coprolites. Scale bar, 5 mm. **b**, FMNH PE10638, *Tullimonstrum* with preserved

tectal cartilages (arrow). Scale bar, 10 mm. **c**, FMNH PE39169, *Tullimonstrum* with liver (arrow) preserved as a diffuse dark circular structure. Inset is a line drawing: red, notochord; brown, liver. Scale bar, 10 mm.

Extended Data Table 1 | Summary of specimen descriptions

Morphological Feature	Number of specimens that could preserve the feature	Number of specimens that do preserve the feature	Percentage
Teeth	75	15	20%
Curved proboscis	729	33	5%
Bent proboscis	729	153	21%
Myomeres in body	1058	505	48%
Myomeres in tail	510	139	27%
Notochord	1058	204	19%
Arcualia	1058	357	34%
Eye bar	1058	785	74%
Eyes	785	446	57%
Caudal fin	510	442	87%

A summary of the specimen data presented in Supplementary Table 1. If the concretion is broken off at the eye bar, and the piece with the proboscis is lost, for example, that specimen does not have the potential to preserve teeth and would not be considered when counting specimens with teeth.

Extended Data Table 2 | Tail aspect ratio for *Tullimonstrum* and extant lampreys

Species	Tail height (H)	Tail surface area (S)	Aspect ratio (A) $A=H^2/S$
<i>Lethenteron appendix</i>			
American brook lamprey	39.01	3106.00	0.49
<i>Ichthyomyzon castaneus</i>			
Chestnut lamprey	50.04	3207.00	0.78
<i>Ichthyomyzon fossor</i>			
Northern brook lamprey	69.07	7973.00	0.60
<i>Petromyzon marinus</i>			
Sea lamprey	105.02	13550.00	0.81
<i>Ichthyomyzon unicuspis</i>			
Silver lamprey	83.05	12722.00	0.54
<i>Ichthyomyzon gagei</i>			
Southern brook lamprey	51.01	3252.00	0.80
<i>Tullimonstrum gregarium</i>			
FMNH PE60343	25.31	491.07	1.30
<i>Tullimonstrum gregarium</i>			
FMNH PE10654	34.67	1063.89	1.13
<i>Tullimonstrum gregarium</i>			
FMNH PE14125	38.98	1503.52	1.01

Lamprey measurements were obtained from pictures from <http://www.seagrant.wisc.edu>. All measurements are in millimetres.

The eyes of *Tullimonstrum* reveal a vertebrate affinity

Thomas Clements¹, Andrei Dolocan², Peter Martin^{3,4}, Mark A. Purnell¹, Jakob Vinther^{3,5} & Sarah E. Gabbott¹

Tullimonstrum gregarium is an iconic soft-bodied fossil from the Carboniferous Mazon Creek Lagerstätte (Illinois, USA)¹. Despite a large number of specimens and distinct anatomy, various analyses over the past five decades have failed to determine the phylogenetic affinities of the ‘Tully monster’, and although it has been allied to such disparate phyla as the Mollusca², Annelida^{3,4} or Chordata⁵, it remains enigmatic^{1–5}. The nature and phylogenetic affinities of *Tullimonstrum* have defied confident systematic placement because none of its preserved anatomy provides unequivocal evidence of homology, without which comparative analysis fails. Here we show that the eyes of *Tullimonstrum* possess ultrastructural details indicating homology with vertebrate eyes. Anatomical analysis using scanning electron microscopy reveals that the eyes of *Tullimonstrum* preserve a retina defined by a thick sheet comprising distinct layers of spheroidal and cylindrical melanosomes. Time-of-flight secondary ion mass spectrometry and multivariate statistics provide further evidence that these microbodies are melanosomes. A range of animals have melanin in their eyes, but the possession of melanosomes of two distinct morphologies arranged in layers, forming retinal pigment epithelium, is a synapomorphy of vertebrates. Our analysis indicates that in addition to evidence of colour patterning⁶, ecology⁷ and thermoregulation⁸, fossil melanosomes can also carry a phylogenetic signal. Identification in *Tullimonstrum* of spheroidal and cylindrical melanosomes forming the remains of retinal pigment epithelium indicates that it is a vertebrate; considering its body parts in this new light suggests it was an anatomically unusual member of total group Vertebrata.

The enigmatic *T. gregarium* from the Carboniferous Mazon Creek Lagerstätte (307 million years ago) is among the world’s most controversial fossils. Familiar to millions of people as the official state fossil of Illinois, reconstructions of the ‘Tully monster’ have graced the sides of U-Haul trailers across the USA. Yet the phylogenetic affinity of *Tullimonstrum* remains unresolved. In contrast with the Cambrian Chengjiang and Burgess Shale biotas, the Mazon Creek preserves fossils that are largely familiar (at least at the level of higher taxa), with *Tullimonstrum* being a notable anomaly in this respect. *Tullimonstrum*, a monotypic taxon known from several hundred specimens, is preserved as stains with some relief within Mazon Creek siderite nodules. Despite the uncertainty about its position in the tree of life, there is a surprisingly high level of agreement about the arrangement and shape of anatomical features (Fig. 1 and Table 1). The anatomical complexity, evident cardinal axes and the bilateral symmetry demonstrates that *Tullimonstrum* is a bilaterian^{1–5}; however, beyond this, it has defied systematic placement. Given the consensus about the shape and anatomical disposition of body parts, this might seem perplexing. But the issue is in fact quite simple: there is little agreement about its affinities because no study has identified unequivocal homologies/synapomorphies upon which to base a solid comparative anatomical interpretation. This is a classic example of how, without the criterion of topological

relations between body parts as a potential falsifier of character hypotheses, testing alternative hypotheses becomes problematical^{9,10}. Different choices of extant anatomical comparator result in radically different hypotheses of homology and affinity for *Tullimonstrum* (Table 1), but evidence to test which hypothesis is correct remains elusive. Where topological data in fossils are equivocal, other homology criteria, normally subordinate to topology, assume greater importance^{9–11}. Here we apply the criterion of the intrinsic properties of body parts (also referred to as ‘special qualities’¹⁰ or ‘correspondence of composition’¹¹) to resolve the phylogenetic placement of *Tullimonstrum*.

One of the defining characters of *Tullimonstrum* is the transverse bar. Associated with this in many specimens is a pair of dark structures that, regardless of the orientation of the fossil, occur at the distal ends of the bar (Figs 1 and 2 and Extended Data Fig. 1). The transverse bar is relatively straight, although it bends forwards or backwards in some specimens³; it is preserved in relief, suggesting a relatively recalcitrant structure, but there is no evidence that it was biomineralized³. Scanning electron microscopy and energy-dispersive X-ray spectroscopy reveal that the dark structures comprise thick, multi-layered masses of tightly packed, micrometre-sized bodies composed of carbonaceous material (Fig. 2). They exhibit two distinct morphologies: highly cylindrical forms with rounded terminations (1.3–2.0 µm long and 0.3–0.4 µm wide), and oblate, almost spherical forms (0.4–0.7 µm diameter). There are at least two layers of bodies, with oblate and cylindrical types showing little intermixing (Fig. 2 and Extended Data Fig. 1). No other anatomy, even that composed of carbon, exhibits this microtexture (Extended Data Fig. 2).

The composition, anatomical localization and fabrics indicate that the cylindrical and oblate bodies are layers of melanosomes; the range of shape and size compares closely with extant and fossilized



Figure 1 | *T. gregarium* fossil from the Mazon Creek Lagerstätte. Optical image (BMRP2014MCP1000) showing typical morphology and spatial relationships between the principal anatomical features: 1, appendage; 2, stylets in terminal structure; 3, transverse bar; 4, distal structures on transverse bar; 5, transverse sigmoidal bands on trunk; 6, extensions to the posterior body. Differing opinions on these anatomical characters in the literature are shown in Table 1. Scale bar, 40 mm.

¹Department of Geology, University of Leicester, Leicester LE1 7RH, UK. ²Texas Materials Institute, The University of Texas at Austin, Austin, Texas 78712, USA. ³School of Earth Sciences, University of Bristol, Bristol BS8 1RJ, UK. ⁴Interface Analysis Centre, HH Wills Physics Laboratory, University of Bristol, Bristol BS8 1TQ, UK. ⁵School of Biological Sciences, University of Bristol, Bristol BS8 1TQ, UK.

Table 1 | Anatomical interpretations and affinity of *T. gregarium* in the literature

Reference	1 Appendage	2 Stylets in terminal structure	3 Transverse bar	4 Distal structures on transverse bar	5 Transverse sigmoidal bands on trunk	6 Extensions to the posterior body	Proposed affinity
1	Jaw apparatus	Stylets/teeth	N/A	Bar organs	Segmentation	Lateral tail fins	Incertae sedis
3	Grasping claw Jaw	Stylets	Sensory organs Otocysts Hydrodynamic stabilizers	Eyes	Segmentation	Lateral tail fins	Nemertea Polychaeta Sipunculidea Arthropoda Echiuroidea
2	Buccal mass	Teeth	Eye stalks	N/A	Segmentation Muscle bands	Lateral tail fins Dorsoventral tail fins	Nemertea Polychaeta Mollusca
5	Grasping claw Jaw buccal mass	Teeth	Paired copulatory organs Setae	N/A	Segmentation Muscle bands	Dorsoventral tail fins	Mollusca <u>Conodonta</u>
4	N/A	N/A	N/A	N/A	Segmented muscles	Caudal appendage	Nemertea Annelida
Proposed herein	Proboscis	Tentative: teeth or dermal denticles?	Eye stalks	Eyes	Possible myomeres?	Dorsoventral tail fins	Non-osteichthyan total group vertebrate; possible total group gnathostome

Anatomical characters correspond with Fig. 1. 'Proposed affinity' lists the range of groups that *Tullimonstrum* has been allied to. Underlined groups indicate the phylogenetic placement favoured in each original study.

melanosomes⁶. To further test this hypothesis we employed time-of-flight secondary ion mass spectrometry (TOF-SIMS) and principal component analysis (PCA) to compare the relative intensity distribution of the melanin-specific peaks originating from fresh, artificially matured, fossil melanin and non-melanin samples (Extended Data Fig. 3).

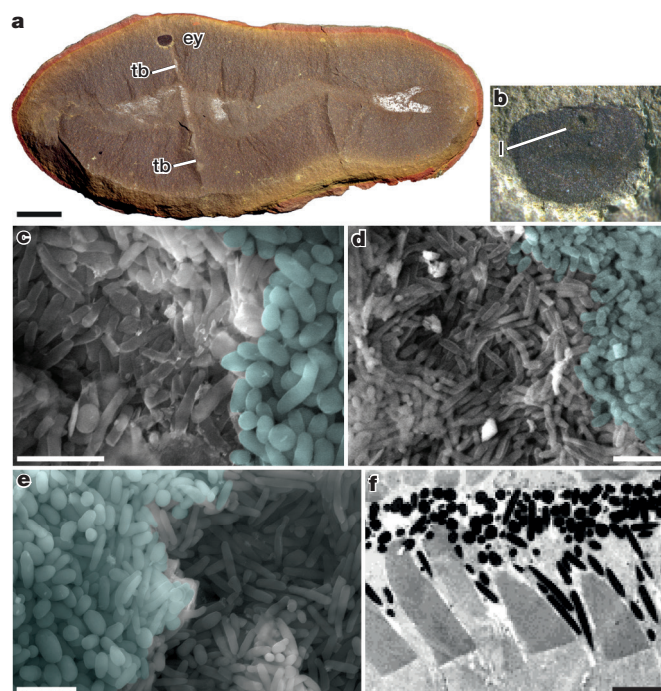


Figure 2 | The ultrastructural details of the eyes of *T. gregarium* from the Mazon Creek Lagerstätte and an extant anchovy (*Coilia nasus*). **a**, *Tullimonstrum* (PE22061) with eye (ey) and transverse bar (tb). **b**, Eye in **a**, with possible lens (l) (see Extended Data Fig. 1). **c–e**, Scanning electron microscope images of melanosomes in *Tullimonstrum* eye (**c**, **d**, PE22061; **e**, PE22126). The boundary between layers is highlighted by blue applied to areas dominated by oblate melanosomes. **f**, Radial cross section (transmission electron microscope) of a larval anchovy retina with oblate and cylindrical melanosomes in the RPE (dark pigment granules). Image used with permission²⁹. Decay-induced collapse of the RPE would result in a fossilized structure with oblate melanosomes overlying cylindrical as seen in **c–e** (see Extended Data Fig. 1), or vice versa, depending on specimen orientation. Scale bars, 10 mm (**a**); 2 μ m (**c–f**).

Spectra from *Tullimonstrum* and pure melanin samples¹² show a similar spectral composition (Fig. 3 and Extended Data Fig. 3). PCA show *Tullimonstrum* data plot among samples of fossil melanin¹² (Fig. 3 and Extended Data Fig. 4), thus providing, in addition to anatomical localization and morphology, independent chemical evidence that the microbodies are melanosomes. An alternative interpretation is that microbodies are the remains of melanin-synthesizing bacteria or fungi. This scenario is unlikely because these microorganisms are not known to colonize decaying bodies, and their distribution in the fossils would require that they localized only to formerly melanin-synthesizing tissues.

Within the Mazon Creek Lagerstätte the only other fossils to possess paired, dark, ovoid structures are the numerous vertebrates (cyclostomes and gnathostomes), and a single putative coleoid¹³. In vertebrates, anatomical landmarks indicate that the dark structures are eyes (see, for example, refs 14–17 and Extended Data Fig. 5). Eyes in basal vertebrates are relatively decay resistant^{18,19} and pigment is one of the most decay resistant features in lampreys^{18,19}. In *Tullimonstrum*, the dark structures are paired, bilaterally disposed and comprise thick, multi-layered masses of melanosomes. Together, these data constitute strong evidence that the dark structures are eyes.

Retinal pigments function as visual photoreceptors or as screening pigments that act to prevent stray light from reaching the photoreceptive cells²⁰. While all metazoans can synthesize melanin, ocular screening pigments are known to vary, and current data indicate that invertebrates chiefly employ ommochromes and pterines²¹. In annelids, molluscs and arthropods these pigments are contained in microbodies that are exclusively spherical or slightly oval, frequently faceted by abutting pigment granules and cell walls. There are a handful of invertebrate groups where melanin has been chemically identified as the screening pigment (planarian flatworms²², cubozoan cnidarians²³ and ascidians²⁴, phaeomelanin in the shell-eyes of chitons²⁵). Significantly, the available ultrastructural data indicate that where these groups employ melanin, their melanosomes are exclusively ovoid (Fig. 4; see also Supplementary Information).

Chordates are unusual among metazoans in that their ocular screening pigments are exclusively melanin²³. In vertebrate eyes, the iris, choroid and retinal pigment epithelium (RPE) all contain melanosomes but the last tissue is distinct in having layers of ovoid and cylindrical melanosomes²⁶. *Tullimonstrum* eyes comprise ovoid and cylindrical melanosomes that occur in distinct layers (that is, not intermixed; Fig. 2 and Extended Data Fig. 1) and we therefore interpret the melanosome layer as the remains of RPE. The possibility that this micro-anatomical complex—melanosomes of the same size and shape, arranged in layers,

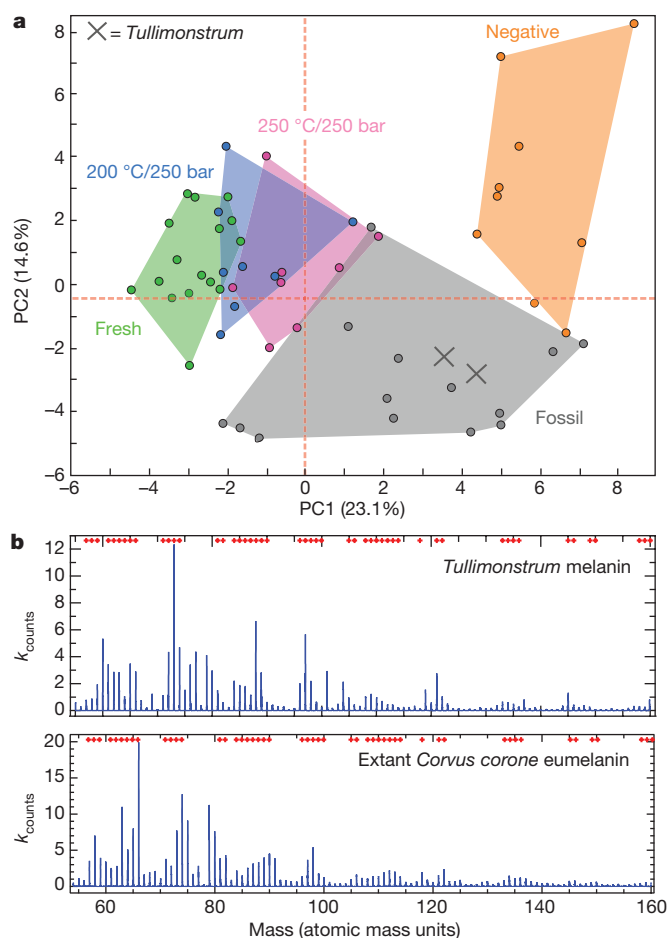


Figure 3 | TOF-SIMS analysis of melanosomes preserved in *T. gregarium*. **a**, PCA of 55 negative secondary ion peaks¹² from fresh, artificially matured (24 h at 200 °C/250 bar and 250 °C/250 bar) and fossil melanin samples as well as a variety of known melanin-negative samples. Two measurements from the eye of *Tullimonstrum* (BMRP2014MCP1000) are marked as 'X'. Two separately acquired spectra from regions of the eye in *Tullimonstrum* reveal a composition (**b**) with similar relative intensity distributions of the melanin-specific peaks (indicated by red crosses) to extant melanin samples (for example, extant crow melanin). However, the PCA analysis indicates that fresh and fossil melanins are quantifiably different. Artificially matured melanins plot closer to fossil samples, suggesting diagenetic alteration of fossil melanin¹². The *Tullimonstrum* spectrum is most similar to that from an Eocene frog eye as well as a lamprey eye from Mazon Creek (see Extended Data Figs 3 and 4 for loadings and details). Red crosses in **b** indicate eumelanin-characteristic fragments.

exclusively in the eye—was convergently acquired by *Tullimonstrum* and vertebrates is non-parsimonious. On the basis of the available evidence from extant animals this character complex is a synapomorphy of vertebrates, and it thus represents an unequivocal phylogenetically informative homology in *Tullimonstrum*.

The homology of RPE in *Tullimonstrum* provides the phylogenetic context for comparative anatomical evaluation. A full analysis, including comparative taphonomy, is beyond the scope of this contribution, but here we consider the main body parts (Fig. 1 and Table 1), particularly diagnostic characters of the Chordata such as the notochord and myomeres.

Of the generally accepted body parts in *Tullimonstrum*, none is readily interpreted as a notochord or a branchial structure. Fossil lamprey and hagfish (that is, non-biomineralized vertebrates) from the Mazon Creek also lack a preserved notochord^{14,15}, suggesting that absence in *Tullimonstrum* is likely to reflect a failure to fossilize rather than an absence from the organism. Similarly, branchial structures of Mazon

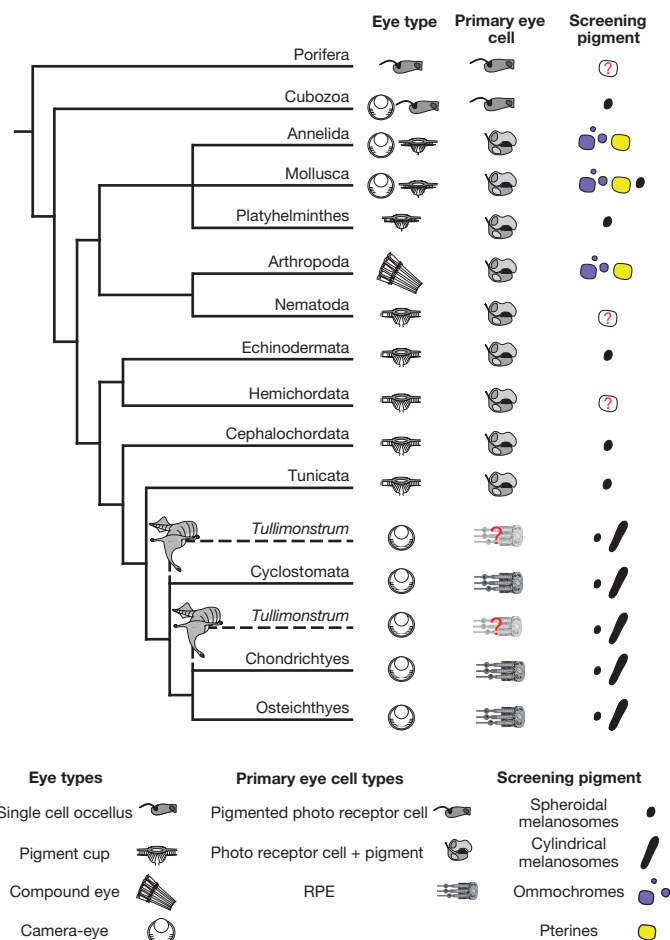


Figure 4 | The phylogenetic distribution of photoreceptor organs, cell architecture and pigment granule chemistry and morphology in animals. Vertebrates are unique among eye-bearing metazoans in having their screening pigment cells and photoreceptors in separate tissues (RPE and the rods and cones layer; although, as seen in Fig. 2f, the RPE layer may project in between the rods and cones). In all other metazoans these cells inter-finger to form a retinal layer, or are combined into a single cell in basal metazoans and some protostomes. Possible positions of *Tullimonstrum* within total group vertebrates are indicated. For further details, see Supplementary Information.

Creek lampreys and hagfish are preserved in a way that indicates they were pigmented in life (S.E.G., unpublished observations); we take their absence from *Tullimonstrum* to indicate that they were unpigmented. V-shaped stains interpreted as myomeres are known from Mazon Creek agnathans (see, for example, ref. 27), and the hypothesis that the transverse sigmoidal bands of the trunk in *Tullimonstrum* represent myomeres or myosepta is possible. The asymmetrical, oblongate posterior fins of *Tullimonstrum* have generally been reconstructed as dorsoventrally flattened³, and this would be unusual in a vertebrate. However, analysis indicates that the tail was laterally flattened in life and that the apparent dorso-ventral flattening in some specimens is a result of post-mortem twisting, evidenced by oblique wrinkles commonly seen in the posterior portion of the body immediately anterior to the tail^{2,5}, also seen in Mazon Creek chondrichthyans¹⁶. So *Tullimonstrum*, when considered through a taphonomic filter, does preserve some features consistent with a vertebrate body plan.

The most perplexing features of *Tullimonstrum* are the proboscis-like anterior, terminating in a claw-like structure, and the transverse bar. The former remains contentious as it is difficult to determine whether the distal end is a buccal mass², a grasping claw³ or a flexible proboscis. Under a vertebrate model, the 'stylets' could represent biomineralized teeth or dermal denticles, and this is consistent with their mouldic

preservation, comparable to biomineralized structures in Mazon Creek gnathostomes¹⁷. If the 'claw' is a buccal mass this might reflect anterior rostralization or posterior displacement of the eye. Perhaps more likely is the interpretation of this flexible rostral extension as a proboscis, similar to that of the Australian ghost shark, *Callorhynchus milii* (Holocephali). The unusual transverse bar we interpret as a stalked eye structure, on the basis of the presence of melanosomes and the remains of RPE. Stalked eyes occur in several animal groups including vertebrates (for example, larvae of several phylogenetically distinct teleost clades possessing eyes borne on stalks, up to one-quarter the length of the body²⁸; the larvae of *Idiacanthus fasciola*²⁸ and *Stylophthalmus paradoxus* resemble *Tullimonstrum* in having markedly stalked eyes and a rostral extension). Stalked eyes with well-developed RPE (and a possible lens, see Fig. 2 and Extended Data Fig. 1) suggests a camera-style eye capable of image formation, meaning that vision in *Tullimonstrum* involved more than simple detection of light direction as is the case in non-vertebrate chordates.

None of the preserved anatomy of *Tullimonstrum* contradicts the hypothesis that it is a vertebrate, and in the absence of any other unequivocal indicators of homology we show that the intrinsic properties of the eye, a character complex indicative of vertebrate RPE, provide compelling evidence that *Tullimonstrum* is a total group vertebrate. A dual-melanosome RPE evolved at some stage along the vertebrate stem and therefore does not constrain how near the base of the vertebrate tree *Tullimonstrum* might sit. However, if dual-melanosome RPE is a synapomorphy of crown vertebrates, and the stylets in the 'claw' prove to be the remains of biomineralized (phosphatized) structures, the affinities of *Tullimonstrum* would lie with total group gnathostomes. Lacking any evidence of a bony skeleton, a placement within Osteichthyes is unlikely; however, without additional diagnostic characters, *Tullimonstrum* cannot currently be assigned to any more delineated clade.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 12 October 2015; accepted 9 March 2016.

Published online 13 April 2016.

- Richardson, E. S. Jr. Wormlike fossil from the Pennsylvanian of Illinois. *Science* **151**, 75–76 (1966).
- Foster, M. in *Mazon Creek Fossils* (ed. Nitecki, M. H.) 269–301 (Academic, 1979).
- Johnson, R. G. & Richardson, E. S. Pennsylvanian invertebrates of the Mazon Creek Area, Illinois: the morphology and affinities of *Tullimonstrum*. *Fieldiana Geol.* **12**, 119–149 (1969).
- Schram, F. in *The Early Evolution of Metazoa and The Significance of Problematic Taxa* (eds Conway-Morris, S. & Simonetta, A. M.) 35–46 (Cambridge Univ. Press, 1991).
- Beall, B. in *The Early Evolution of Metazoa and The Significance of Problematic Taxa* (eds Conway-Morris, S. & Simonetta, A. M.) 271–286 (Cambridge Univ. Press, 1991).
- Vinther, J., Briggs, D. E., Prum, R. O. & Saranathan, V. The colour of fossil feathers. *Biol. Lett.* **4**, 522–525 (2008).
- Clarke, J. A. et al. Fossil evidence for evolution of the shape and color of penguin feathers. *Science* **330**, 954–957 (2010).
- Lindgren, J. et al. Skin pigmentation provides evidence of convergent melanism in extinct marine reptiles. *Nature* **506**, 484–488 (2014).
- Donoghue, P. C. & Purnell, M. A. Distinguishing heat from light in debate over controversial fossils. *BioEssays* **31**, 178–189 (2009).
- Rieppel, O. & Kearney, M. Similarity. *Biol. J. Linn. Soc.* **75**, 59–82 (2002).

- Ruppert, E. E. Key characters uniting hemichordates and chordates: homologies or homoplasies? *Can. J. Zool.* **83**, 8–23 (2005).
- Colleary, C. et al. Chemical, experimental, and morphological evidence for diagenetically altered melanin in exceptionally preserved fossils. *Proc. Natl Acad. Sci. USA* **112**, 12592–12597 (2015).
- Kluessendorf, J. & Doyle, P. *Pohlsepia mazonensis*, an early 'octopus' from the Carboniferous of Illinois, USA. *Palaeontology* **43**, 919–926 (2000).
- Bardack, D. First fossil hagfish (Myxinoidea): a record from the Pennsylvanian of Illinois. *Science* **254**, 701–703 (1991).
- Bardack, D. & Zangerl, R. First fossil lamprey: a record from the Pennsylvanian of Illinois. *Science* **162**, 1265–1267 (1968).
- Sallan, L. C. & Coates, M. I. The long-rostrum elasmobranch *Bandringa zangeri*, 1969, and taphonomy within a Carboniferous shark nursery. *J. Vertebr. Paleontol.* **34**, 22–33 (2014).
- Shabica, C. W. & Hay, A. *Richardson's Guide to The Fossil Fauna of Mazon Creek* (eds Shabica, C. W. & Hay, A. H.) (Northeastern Illinois Univ., 1997).
- Sansom, R. S., Gabbott, S. E. & Purnell, M. A. Decay of vertebrate characters in hagfish and lamprey (Cyclostomata) and the implications for the vertebrate fossil record. *Proc. R. Soc. B* **278**, 1150–1157 (2011).
- Sansom, R. S., Gabbott, S. E. & Purnell, M. A. Atlas of vertebrate decay: a visual and taphonomic guide to fossil interpretation. *Palaeontology* **56**, 457–474 (2013).
- Fein, A. & Szuts, E. Z. *Photoreceptors, Their Role in Vision* Vol. 5 (Cambridge Univ. Press, 1982).
- Vopalensky, P. & Kozmik, Z. Eye evolution: common use and independent recruitment of genetic components. *Phil. Trans. R. Soc. B* **364**, 2819–2832 (2009).
- Hase, S. et al. Characterization of the pigment produced by the planarian, *Dugesia ryukyuensis*. *Pigment Cell Res.* **19**, 248–249 (2006).
- Kozmik, Z. et al. Assembly of the cnidarian camera-type eye from vertebrate-like components. *Proc. Natl Acad. Sci. USA* **105**, 8989–8993 (2008).
- Sato, S. & Yamamoto, H. Development of pigment cells in the brain of ascidian tadpole larvae: insights into the origins of vertebrate pigment cells. *Pigment Cell Res.* **14**, 428–436 (2001).
- Speiser, D. I., DeMartini, D. G. & Oakley, T. H. The shell-eyes of the chiton *Acanthopleura granulata* (Mollusca, Polyplacophora) use pheomelanin as a screening pigment. *J. Nat. Hist.* **48**, 2899–2911 (2014).
- Liu, Y. et al. Comparisons of the structural and chemical properties of melanosomes isolated from retinal pigment epithelium, iris and choroid of newborn and mature bovine eyes. *Photochem. Photobiol.* **81**, 510–516 (2005).
- Bardack, D. & Richardson, E. Jr. New agnathous fishes from the Pennsylvanian of Illinois. *Fieldiana Geol.* **33**, 489–510 (1977).
- Weihs, D. & Moser, H. Stalked eyes as an adaptation towards more efficient foraging in marine fish larvae. *Bull. Mar. Sci.* **31**, 31–36 (1981).
- Haacke, C., Hess, M., Melzer, R. R., Gebhart, H. & Smola, U. Fine structure and development of the retina of the grenadier anchovy *Coilia nasus* (Engraulidae, Clupeiformes). *J. Morphol.* **248**, 41–55 (2001).

Supplementary Information is available in the online version of the paper.

Acknowledgements W. Simpson, P. Mayer, S. Williams, D. Rudkin and K. Seymour are thanked for specimen access and loans. Funding was through a Natural Environment Research Council studentship P14DF19 (to T.C.) and grant NE/K004557/1 (to M.A.P. and S.E.G.). We also acknowledge the National Science Foundation grant DMR-0923096 used to purchase the TOF-SIMS instrument at Texas Materials Institute, UTA. D. Murdock, C. Nedza, S. Wentges and A. Clements are thanked for proofreading. S. Furzeland is thanked for scanning electron microscope optimization. P. Smith is thanked for Adobe Illustrator tutorials.

Author Contributions S.E.G. and M.A.P. conceived the research programme of which this work is part. S.E.G., J.V. and T.C. designed and performed research. S.E.G., T.C., J.V., M.A.P. and A.D. wrote the manuscript. A.D. and J.V. undertook TOF-SIMS analyses and interpretation. J.V., A.D. and M.A.P. conducted PCA. S.E.G., T.C., J.V. and P.M. operated and optimized the scanning electron microscope.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.E.G. (sg21@le.ac.uk) or J.V. (jakob.vinther@bristol.ac.uk).

METHODS

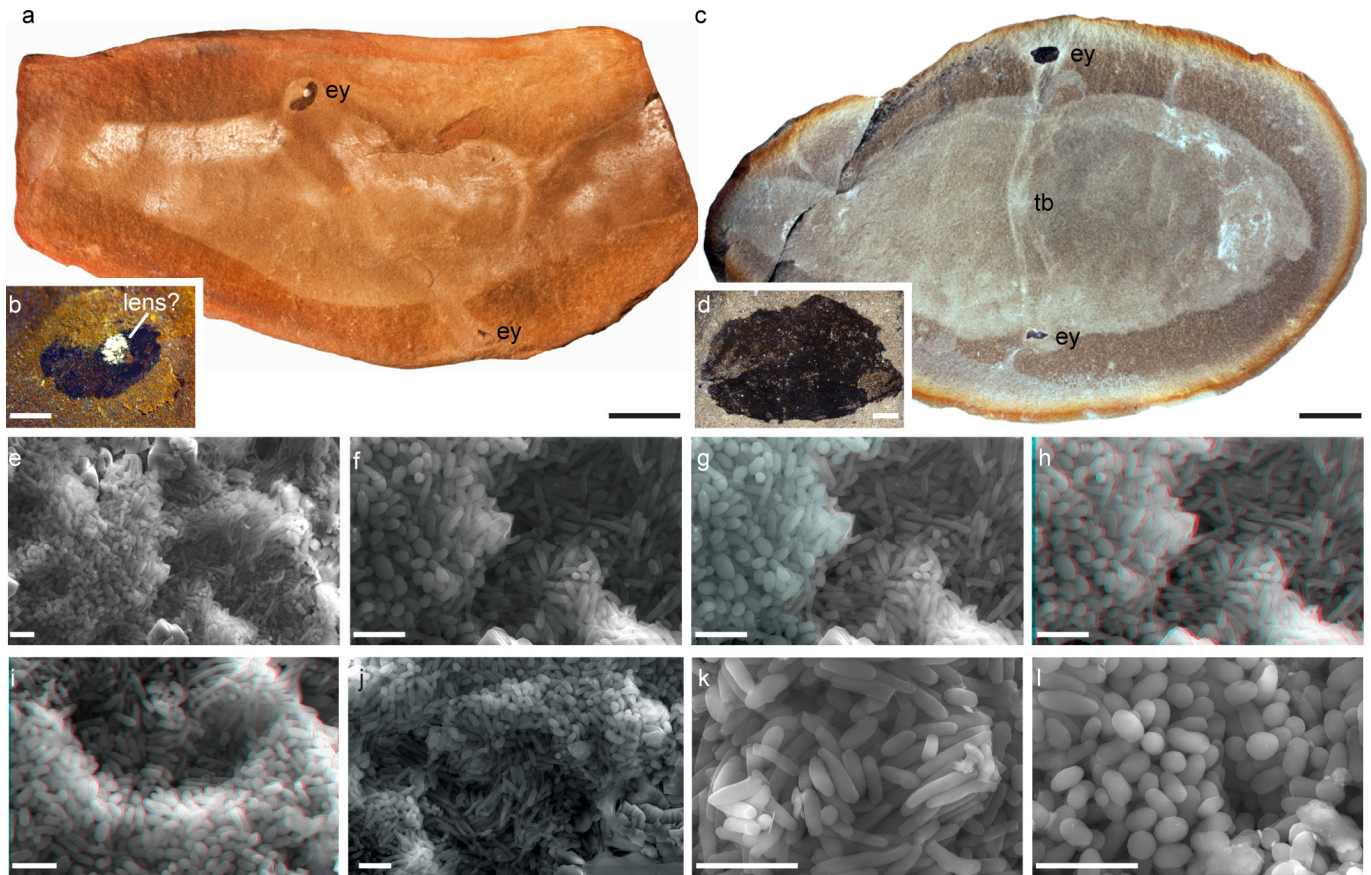
No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

As part of a larger study on pigment preservation and taphonomy in the Mazon Creek, we investigated the dark elliptical patches at the terminations of the transverse bar in 12 specimens of *T. gregarium* from the Burpee Museum of Natural History, Illinois, and the Field Museum of Natural History, Illinois. We analysed textural and compositional data using a Hitachi S-3600N and Zeiss Sigma Environmental scanning electron microscope with an energy-dispersive X-ray spectroscopy system. Partial pressure was 20–30 Pa, working distance was between 9 and 12 mm, with an operating voltage of 15 kV. Specimens were uncoated. Specimens were optically imaged, using a Canon EOS 5D SLR camera and a Leica M205 C stereo microscope.

For TOF-SIMS analysis, one of the eyes in MCPX27C5369 (Burpee Museum of Natural History) was used. The specimen was placed in a TOF.SIMS 5 (ION-TOF, 2010) and secondary ion spectra were collected using a polyatomic analysis beam (Bi_3^+ , 30 keV, 0.9 pA sample current) to increase the yield of organic fragments, as previously employed in ref. 12. Two $500\text{ }\mu\text{m} \times 500\text{ }\mu\text{m}$ areas were analysed in negative polarity with a resolution of $512\text{ pixels} \times 512\text{ pixels}$: one area included the eye and adjacent matrix, another region was selected within the main body of the eye. The acquired maps from within the eye showed no significant effect of topography and was analysed without further processing,

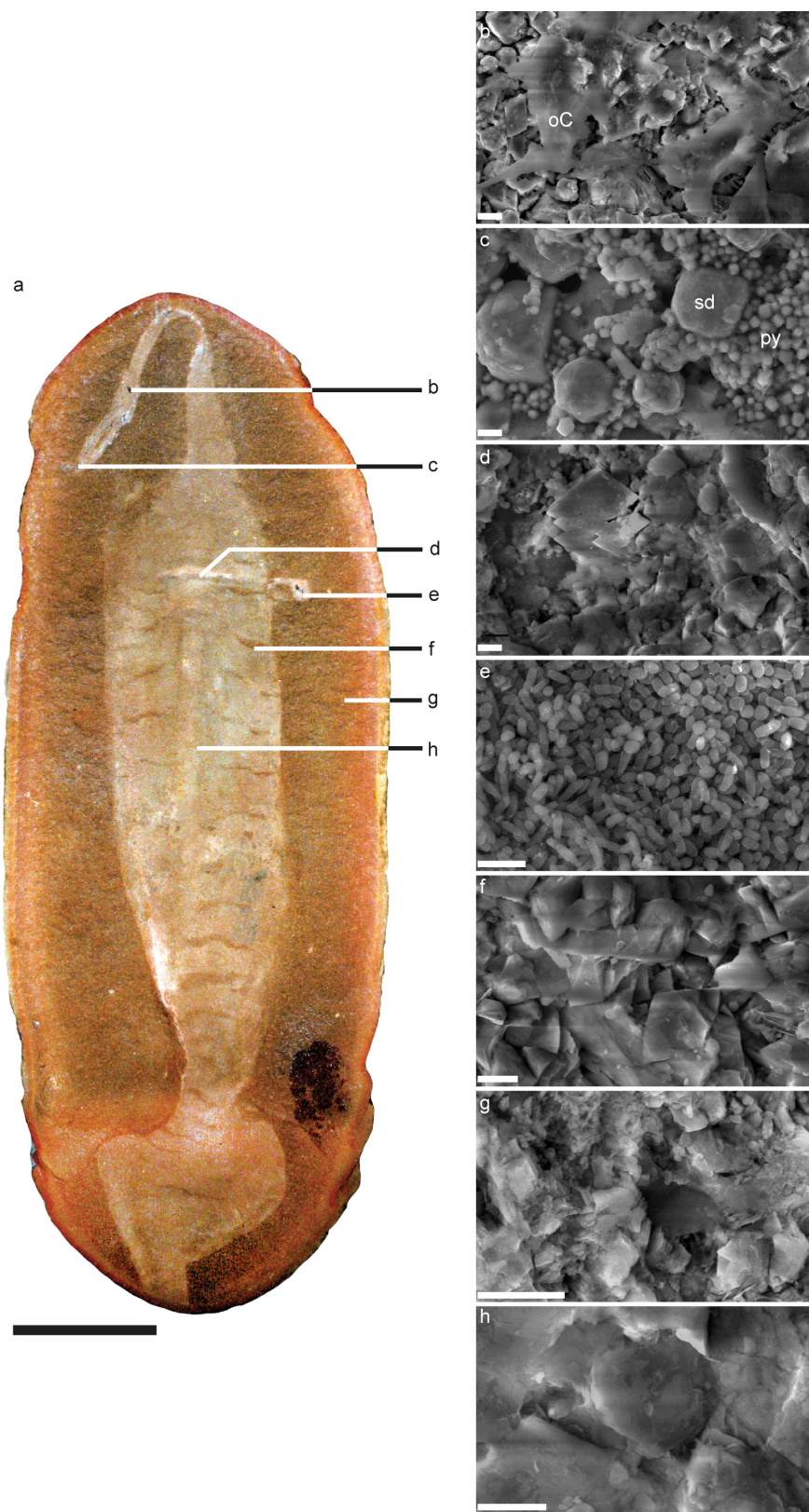
while a region of interest was chosen from the sampled area of the eye and sediment to minimize topographic-related artefacts. All spectra were mass calibrated using the polyatomic fragment series of carbon (C^- , C_2^- , C_3^- , C_4^- , C_5^- , C_7^- , C_8^- , C_9^- , C_{10}^-). The total count intensities of 55 select secondary ion peaks representative for melanin were used for PCA in conjunction with a previously collected data set¹² of artificially matured melanin. Before PCA, each melanin-specific spectrum was normalized to its total intensity, the resulting data set was mean-centred and then standard-deviation-normalized across all samples for each composing mass³⁰. The last process ensured that each melanin-specific peak was given the same weight in the PCA. The *Tullimonstrum* spectra are shown alongside extant reference melanin samples: black (eu)melanosomes from a glossy carrion crow (*Corvus corone*; Fig. 3b and Extended Data Fig. 3), reddish brown domestic chicken (*Gallus gallus*; Extended Data Fig. 3) and representative fossil samples (Jurassic ink sac and Eocene frog eye; Extended Data Fig. 3). Spatial mapping of the melanin-characteristic fragments of the melanosomes within the eye region show a clear separation at micrometre level between the cement (Extended Data Fig. 6e–h) and sediment (Extended Data Fig. 6i–l, u–x), while certain inorganic ions, attributed to calcium phosphates, occur associated with the melanosomes (Extended Data Fig. 6q–t).

30. Wagner, M. S., Graham, D. J., Ratner, B. D. & Castner, D. G. Maximizing information obtained from secondary ion mass spectra of organic thin films using multivariate analysis. *Surf. Sci.* **570**, 78–97 (2004).



Extended Data Figure 1 | Details of *T. gregarium* eyes. **a**, Complete specimen (MCPX27C5369, Burpee Museum of Natural History) with eyes (ey). Scale bar: 10 mm. **b**, Close-up of the uppermost eye in **a**, showing dark carbonaceous material and an approximately centrally positioned white circular area with high relief. The white mineral is kaolinite. This is similar to the eye in *Bandringa* (Extended Data Fig. 6) and the white infilling of kaolinite may be indicative of a lens¹⁶. Scale bar, 1 mm. **c–l**, Specimen PE22126 (Field Museum of Natural History). **c**, Complete specimen in nodule showing clearly defined eyes (ey) and transverse bar (tb). Scale bar, 10 mm. **d**, The uppermost eye in **c**. Scale bar: 1 mm.

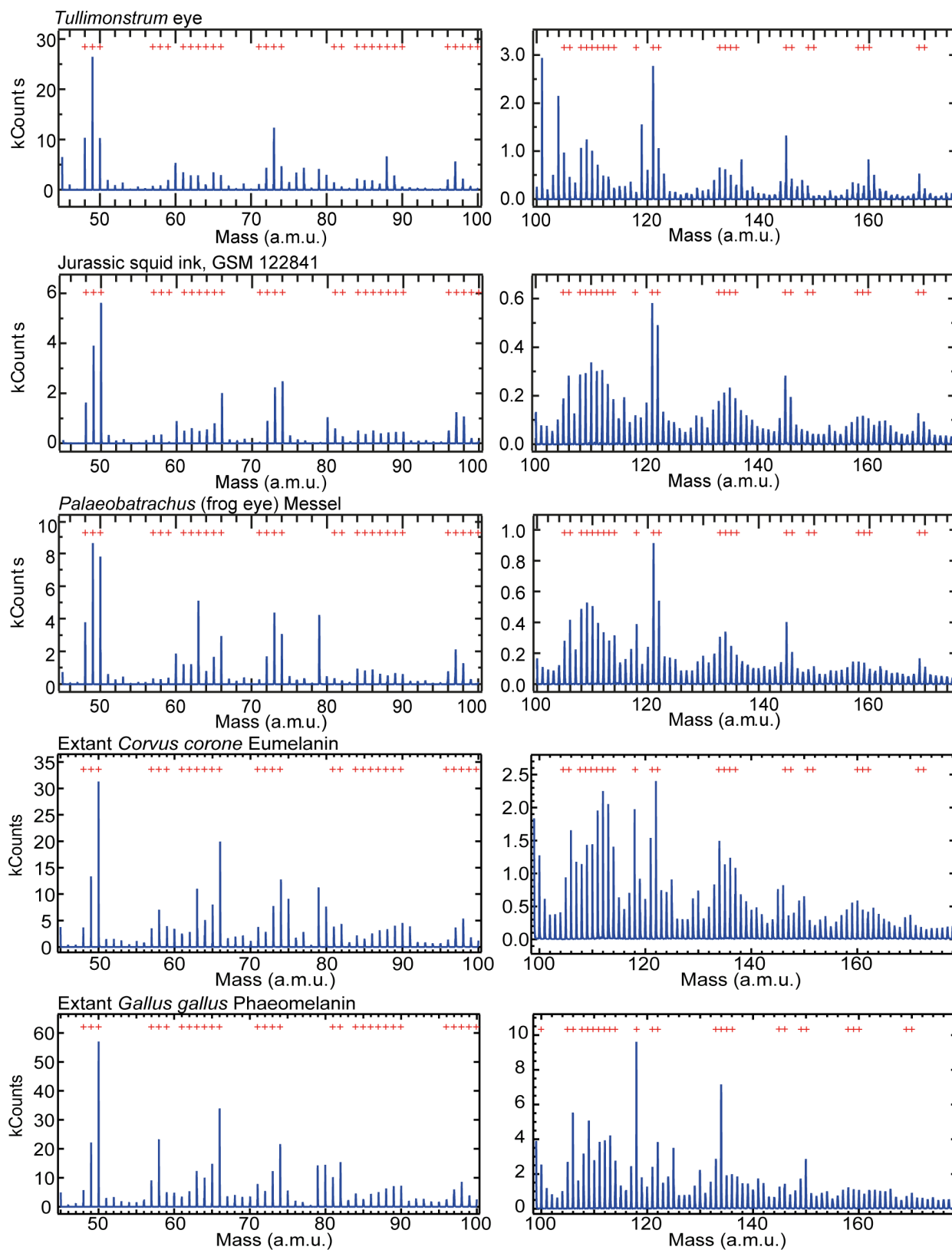
e–l, Scanning electron microscope images of the eye ultrastructure. **e**, Oblate melanosomes on the left-hand side and underlying cylindrical melanosomes on the right-hand side. **f–h**, Higher magnification images of the centre of **e**; in **g**, oblate melanosomes are highlighted in blue, and **h** is an anaglyph (three-dimensional) image of the same field of view as **f** and **g** (see also Fig. 2). **i**, Anaglyph (three-dimensional) image showing oblate melanosomes overlying cylindrical melanosomes. **j**, Oblate and cylindrical melanosomes in distinct layers. **k, l**, The cylindrical and oblate melanosome morphologies, respectively. Scale bars, 2 μ m.



Extended Data Figure 2 | *Tullimonstrum* (BMRP2014MCP1000) with scanning electron microscope images of anatomical features.

a, Complete specimen (anterior at top) with scanning electron microscope images showing the mode of preservation of the anatomy and that only the eyes contain melanosomes. Scale bar, 10 mm. **b**, Proboscis with small, dark, organic carbon patch (oC) which has a smooth texture. **c**, Distal

portion of the proboscis 'claw' showing pyrite crystals and framboids. **d**, Eye bar containing siderite and clay minerals. **e**, Eye showing melanosome texture. **f**, Dark transverse banding (possible myomeres) containing mainly siderite. **g**, The nodule matrix: siderite and detrital clay minerals. **h**, Main trunk: siderite and clay minerals. oC, organic carbon; sd, siderite; py, pyrite. Scale bars, 2 µm.



Extended Data Figure 3 | Negative ion TOF-SIMS spectra in the 45–100 and 100–175 atomic mass unit range. Spectra for comparison with the *Tullimonstrum* eye melanosomes (BMRP2014MCP1000) are from an Eocene frog eye (Messel Lagerstätte), Jurassic ink sac from Lyme Regis (UK), extant glossy black carrion crow (*C. corone*) and a reddish

brown domestic chicken (*G. gallus*). Comparative spectra are from ref. 12. Negative ion TOF-SIMS spectra in the 45–100 atomic mass unit range are shown in the left column and 100–175 atomic mass unit range in the right column. Melanin specific peaks are indicated by red crosses.

b Unmatured samples

1. *Gallus gallus*, phaeomelanin
2. *Corvus corone*, eumelanin
3. *Troglodytes aedon*, brown melanin
4. *Gallus gallus*, eumelanin
5. *Junco hyemalis*, grey feather melanin
6. *Dumetella carolinensis*, grey feather melanin
7. *Anas platyrhynchos*, brown feather melanin
8. *Columba livia*, grey feather melanin
9. *Meleagris gallopavo*, iridescent feather (eumelanin)
10. *Columba livia*, proximal grey feather melanin
11. *Columba livia*, distal black feather melanin
12. *Pelophylax kl. esculentus*, liver melanosomes
13. *Pelophylax kl. esculentus*, eye melanosomes
14. *Pelophylax kl. esculentus*, eye melanosomes
15. *Pica pica*, iridescent feather melanin
16. *Sepia officinalis*, cephalopod ink eumelanin

Matured samples, 24 hours, 200°C/250 bar

17. *Gallus gallus*, phaeomelanin
18. *Corvus corone*, glossy feather eumelanin
19. *Gallus gallus*, black feather eumelanin
20. *Junco hyemalis*, grey feather melanin
21. *Dumetella carolinensis*, grey feather melanin
22. *Anas platyrhynchos*, brown feather melanin
23. *Columba livia*, grey feather melanin
24. *Meleagris gallopavo*, iridescent feather (eumelanin)

Matured samples, 24 hours, 250°C/250 bar

25. *Gallus gallus*, phaeomelanin
26. *Corvus corone*, glossy feather eumelanin
27. *Troglodytes aedon*, brown melanin
28. *Gallus gallus*, eumelanin
29. *Junco hyemalis*, grey feather melanin
30. *Anas platyrhynchos*, brown feather melanin
31. *Columba livia*, grey feather melanin
32. *Meleagris gallopavo*, iridescent feather (eumelanin)

Fossil melanin samples

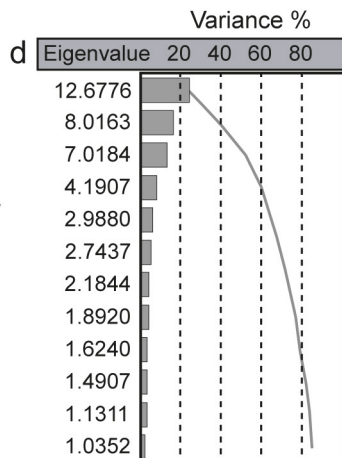
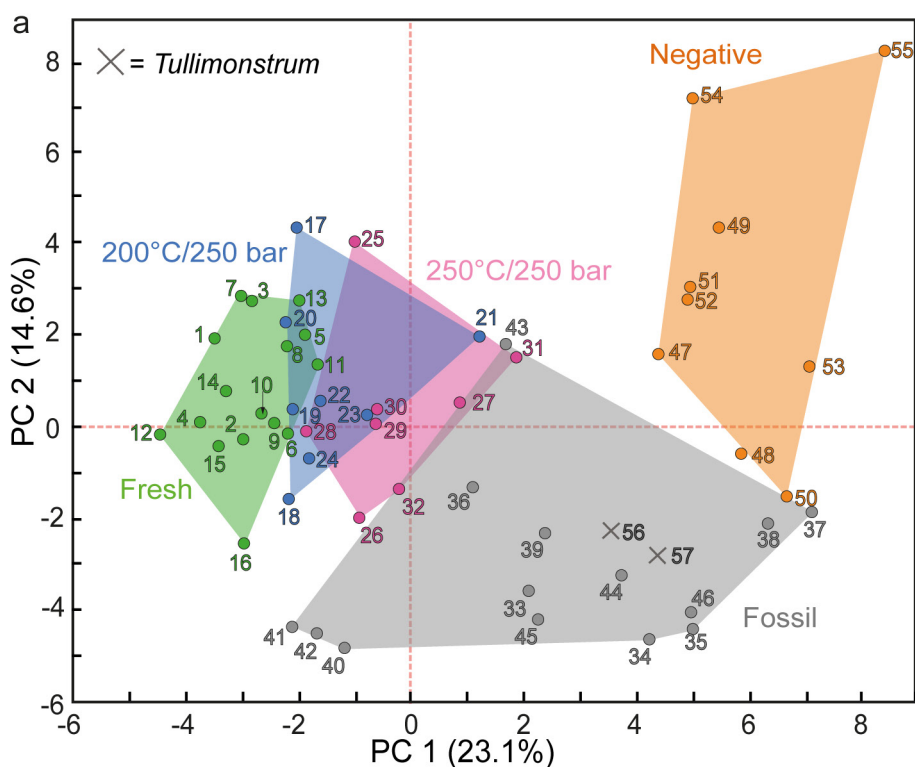
33. Clawed African frog, *Pipidae*, Mush Valley, Miocene
34. Clawed African frog, *Pipidae*, Mush Valley, Miocene
35. Undetermined bird, Fur Formation, Lower Eocene
36. Iridescent feather, Messel, Eocene
37. Hassianxeris, Messel, Eocene
38. *Palaeochiropteryx*, Messel, Eocene
39. Tadpole, *Pelobates*, Enspel, Oligocene
40. Octopus, *Keuppia*, Hakel, Cretaceous
41. Stem octopod, *Glyphiteuthis*, Hakel, Cretaceous
42. Indet. Cephalopod ink sac, Lyme Regis, Lower Jurassic
43. *Messelornis*, wing covert feathers, Messel, Eocene
44. Frog eye, *Palaeobatrachus*, Messel, Eocene
45. Frog skin, *Palaeobatrachus*, Messel, Eocene
46. Lamprey eye, *Mayomyzon*, Mazon Creek, Carboniferous

Non melanin controls

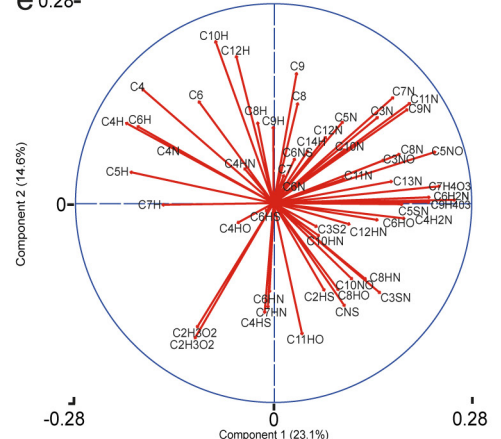
47. Mazon creek nodule sediment, Carboniferous
48. Sediment, Messel, Eocene
49. Oak leaf, Recent
50. Angiosperm leaf, Mush valley, Miocene
51. Organic sediment, Mush valley, Miocene
52. Angiosperm leaf, Stone Rose Quarry, Eocene
53. Sequoia, Stone Rose Quarry, Eocene
54. Bakers yeast, Recent
55. Carbon tape

Tully monsters

56. *Tullimonstrum gregarium*, Mazon Creek, Carboniferous
57. *Tullimonstrum gregarium*, Mazon Creek, Carboniferous

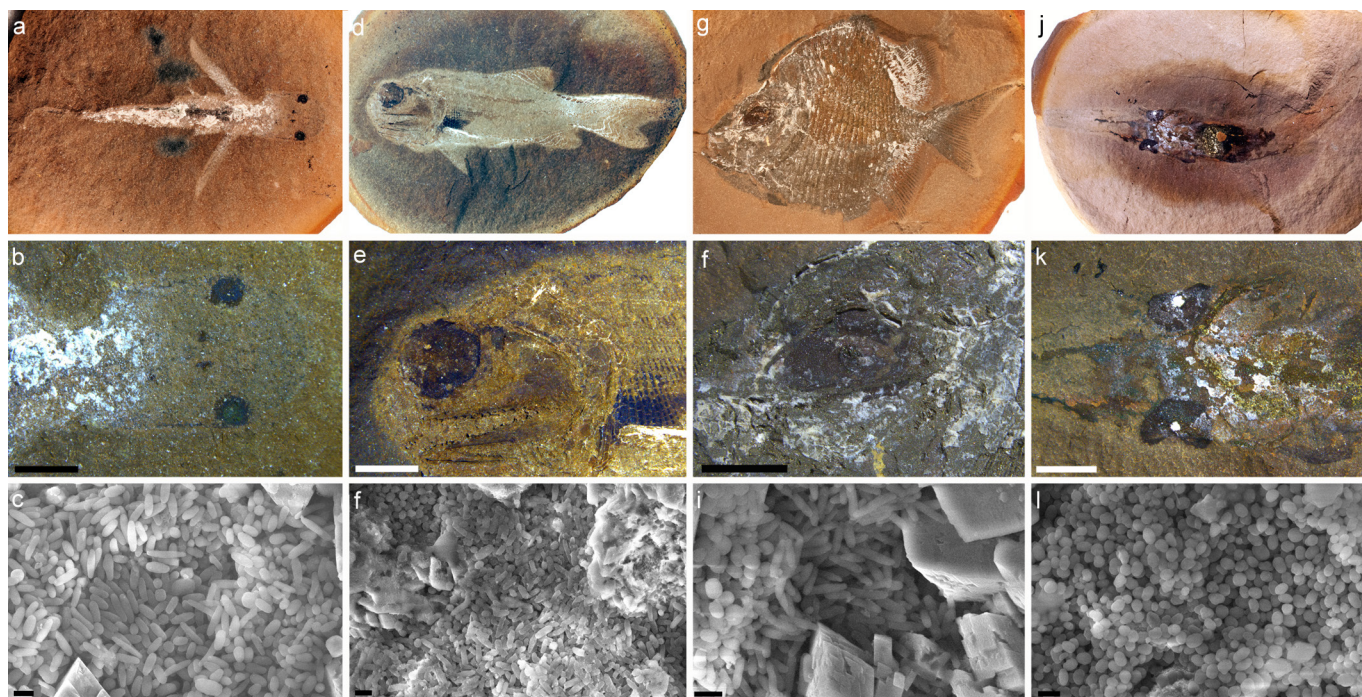


e 0.28-



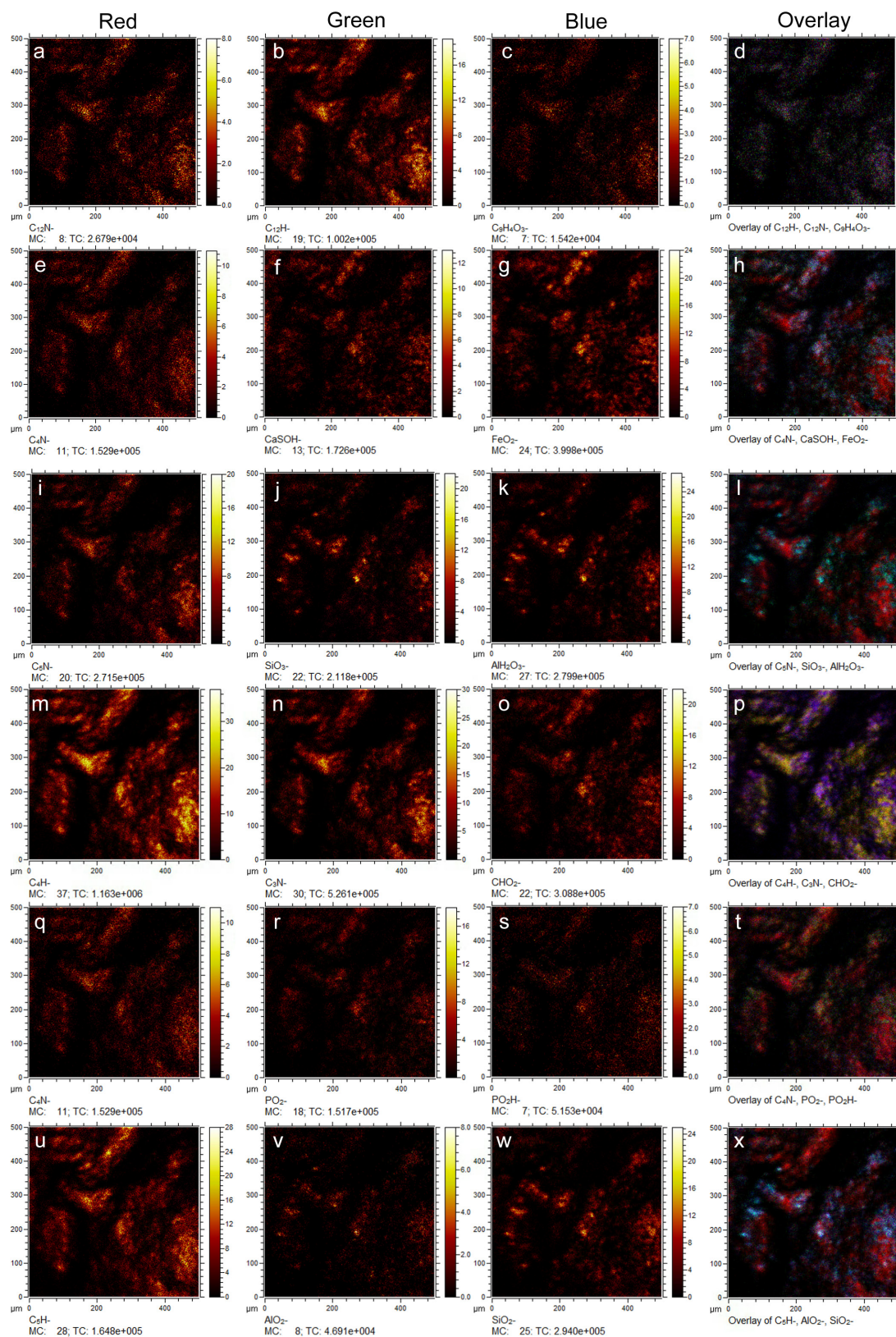
Extended Data Figure 4 | PCA of TOF-SIMS spectra. a, PCA plot of 55 negative secondary ion peaks¹² from fresh, artificially matured (24 h at 200°C/250 bar and 250°C/250 bar) and fossil melanin samples as well as a variety of melanin-negative samples and *Tullimonstrum* eye (all listed in b). The two separately acquired spectra from regions of the eye in *Tullimonstrum* have similar relative intensity distributions of the melanin-specific peaks to other fossil melanosome samples, plotting near an Eocene frog eye and a lamprey eye from Mazon Creek. c, Eigenvector values for principal components 1 and 2. d, Eigenvalues for the first 12 principal

components and the percentage of the variation accounted for by each. e, Loading plot showing the relative factor loadings onto PC axes 1 and 2. Fragments such as C_nNH- , C_nNO- , C_nNS- , C_nSH- and C_nOH- are mostly responsible for the separation of fossil melanin in the PCA space, whereas fragments such as C_n- and C_nH- separate the fresh melanin. This indicates both the chemical degradation (loss of carbon, nitrogen, sulfur and water) and structural degradation (weaker molecular bonding) of melanin during the fossilization process.



Extended Data Figure 5 | Gnathostomes with dark eyes and eye ultrastructure from the Mazon Creek Lagerstätte. a–c, *Esconichthys apopyris* (PF9831), a putative larval lungfish¹⁷. d–f, *Elonichthys peltigerus* (ROM56794). g–i, *Platysomus circularis* (PF7333). j–l, *Bandringa rayi* (ROM56789). Note the white centrally positioned circular feature in both eyes. The white mineral is kaolinite, which most probably reflects the

position of the lens¹⁶. c, f, i, l, Backscattered scanning electron microscope images of the eyes from each corresponding fossil. Melanosomes of cylindrical and oblate morphologies are found in *Esconichthys*, *Elonichthys* and *Platysomus*; in *Bandringa*, only oblate melanosomes occur. Scale bars, 5 mm (b, e, h, k); 1 μ m (c, f, i, l).



Extended Data Figure 6 | TOF-SIMS intensity maps from eye region in *Tullimonstrum*, showing relative distribution of ions derived from melanin relative inorganic ions from the matrix. False-colour chemical mapping of the spatial distribution of several melanin-specific secondary ion fragments (a, e, i, m, q, u) compared with the maps of melanin-characteristic ions (b, c, n, o) and inorganic ions derived from the sediment (SiO₃⁻, j; Al(Hn)O₃ⁿ⁻, k, v) and the concretion cements (FeO₂⁻, g; CaSOH⁻, f), which map distinctly from the melanin ions or co-occur with melanin (PO₂⁻, r; PO₂H⁻, s). The secondary ion CHO₂⁻

is likely from carboxyl groups (o) and is a known constituent of melanin. It exhibits only a moderate overlap with melanin markers (p), which could be attributed to different diagenetic alterations of the melanin or some difference in composition. The right-hand column maps are composites of the tentatively assigned secondary ions in their respective row (that is, d is a composite of a–c). The distribution of inorganic and organic ions shows that the melanin and matrix ions are distinct contributions to the TOF-SIMS spectrum.

In situ imaging reveals the biomass of giant protists in the global ocean

Tristan Biard^{1,2}, Lars Stemmann², Marc Picheral², Nicolas Mayot², Pieter Vandromme³, Helena Hauss³, Gabriel Gorsky², Lionel Guidi², Rainer Kiko³ & Fabrice Not¹

Planktonic organisms play crucial roles in oceanic food webs and global biogeochemical cycles^{1,2}. Most of our knowledge about the ecological impact of large zooplankton stems from research on abundant and robust crustaceans, and in particular copepods^{3,4}. A number of the other organisms that comprise planktonic communities are fragile, and therefore hard to sample and quantify, meaning that their abundances and effects on oceanic ecosystems are poorly understood. Here, using data from a worldwide *in situ* imaging survey of plankton larger than 600 μm , we show that a substantial part of the biomass of this size fraction consists of giant protists belonging to the Rhizaria, a super-group of mostly fragile unicellular marine organisms that includes the taxa Phaeodaria and Radiolaria (for example, orders Collodaria and Acantharia). Globally, we estimate that rhizarians in the top 200 m of world oceans represent a standing stock of 0.089 Pg carbon, equivalent to 5.2% of the total oceanic biota carbon reservoir⁵. In the vast oligotrophic intertropical open oceans, rhizarian biomass is estimated to be equivalent to that of all other mesozooplankton (plankton in the size range 0.2–20 mm). The photosymbiotic association of many rhizarians with microalgae may be an important factor in explaining their distribution. The previously overlooked importance of these giant protists across the widest ecosystem on the planet⁶ changes our understanding of marine planktonic ecosystems.

Oceanic ecosystems are inhabited by a variety of planktonic organisms spanning a wide size range, from nanometres (viruses) to metres

(for example, certain jellyfish). By feeding on small plankton, large zooplankton link primary production to higher trophic levels through the marine food web⁷ and affect carbon export and remineralization to deep oceans by producing fast-sinking particles (fecal pellets and dead bodies)⁸. Most of our knowledge of large zooplankton is based on studies of crustacea such as copepods and euphausiids^{3,4} that are abundant, important for the function of planktonic ecosystems, robust and relatively easy to collect with standard methods such as plankton net tows. As a result, the zooplankton compartment in ecosystem and biogeochemical models is often exclusively represented by the physiological characteristics of copepods⁹.

In contrast, the biology and ecology of planktonic Rhizaria, one of the main eukaryotic super-kingdoms, has been largely unexplored¹⁰. Rhizarians include small unicellular organisms such as Chlorarachniophyta and heterotrophic Cercozoa along with a wealth of larger cells, ranging in size from a few hundred micrometres to several centimetres and belonging to taxonomic groups such as the Radiolaria, Foraminifera and Phaeodaria. These giant (compared to the size of the vast majority of single-celled plankton) protists are predators, but some species are mixotrophs, hosting obligate intracellular microalgal symbionts (photosymbionts)¹¹. Most rhizarians produce mineral skeletons of calcium carbonate (Foraminifera) or silicate (polycystine radiolarians) that are often well preserved in marine sediments, making this group a focus for the development of paleoproxies¹². Others, such as Phaeodaria, Collodaria and Acantharia, possess more delicate

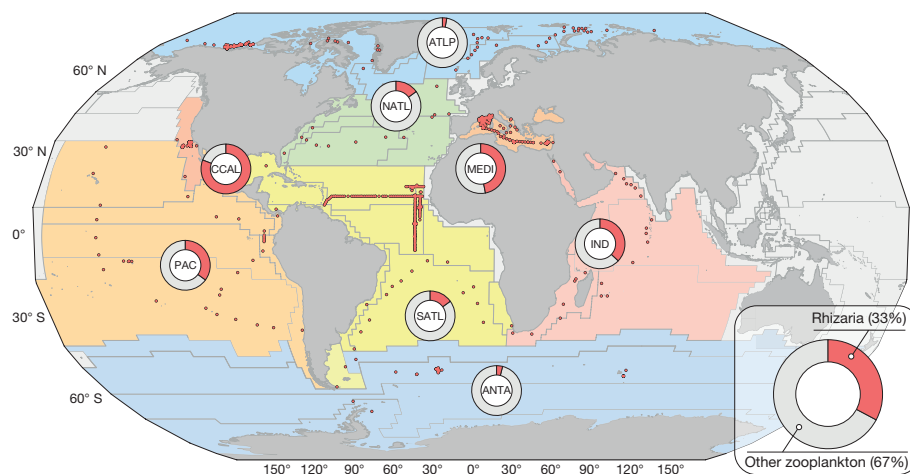


Figure 1 | Worldwide contribution of giant Rhizaria to zooplankton communities (>600 μm) in the top 500 m of the water column.

Underwater Vision Profiler sampling stations are represented by red dots (694 stations; Extended Data Table 1). Relative contributions of the depth-integrated abundances are shown for the Rhizaria (red) and other

zooplankton (grey) as seen and quantified by UVP5. Bottom right panel, global average contribution for each group considered. Contributions are geographically divided according to Longhurst's Biomes and Provinces³⁰ (numerical values are shown in Extended Data Table 2a). Map made with Natural Earth data (<http://www.naturalearthdata.com>).

¹Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire Adaptation et Diversité en Milieu Marin UMR7144, Station Biologique de Roscoff, 29688 Roscoff, France. ²Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire d'Océanographie de Villefranche (LOV) UMR7093, Observatoire Océanologique, 06230 Villefranche-sur-Mer, France. ³GEOMAR Helmholtz Centre for Ocean Research Kiel, Wischhofstrasse 1–3, 24148 Kiel, Germany.

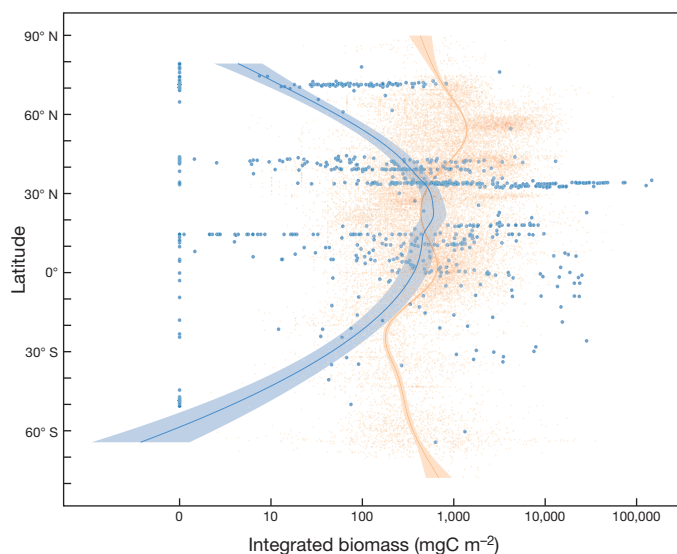


Figure 2 | Latitudinal distribution of depth-integrated biomass (0–200 m depth) of Rhizaria (blue, *in situ* optical assessment, this study; 848 sampling stations) and mesozooplankton (orange, plankton net-based assessments³¹; 26,918 samples). Loess regressions with polynomial fitting were computed to illustrate the latitudinal patterns. Shaded areas represent 95% confidence intervals. Biomass is plotted on a logarithmic scale.

skeletons and are not preserved in marine sediment records^{13,14}. A number of studies, ranging from sediment trap to environmental molecular surveys^{15–17}, have suggested that the Rhizaria are important for present-day oceanic ecosystems. Recent qualitative results from the *Tara* Oceans expedition demonstrated that the Collodaria, which are mainly large colonial rhizarians, are important components of plankton community structure¹⁸ and are significantly correlated with downward fluxes of carbon¹⁹. However, unlike crustacean plankton, which can be easily collected, delicate rhizarians are severely damaged by plankton nets, and other rhizarians, such as Acantharia, eventually dissolve upon preservation in regular fixatives (for example, formalin). Rhizaria are therefore inconsistently sampled^{13,14,20,21} and their global distribution and role in the ecosystem are not well understood. Although they are known to be abundant in specific areas of the oceans^{14,20,22}, their contribution to plankton communities has never been assessed on a global scale.

Using a non-destructive *in situ* imaging system (Underwater Vision Profiler; UVP5)²³, we quantified the respective contributions of Rhizaria and other zooplankton larger than 600 μm (meso- and macro-zooplankton, excluding smaller components of the plankton community) in a variety of pelagic ecosystems (Fig. 1, Extended Data Table 1 and Extended Data Fig. 1). Worldwide, in the upper 500 m of the water column, rhizarians comprised on average 33% of zooplankton observed (Fig. 1). Giant Rhizaria were more abundant in the large inter-tropical oceanic basins, the Mediterranean Sea, and the coastal upwelling off California, where they represented on average 35%, 47% and 81%, respectively, of zooplankton observed (Extended Data Table 2).

When converted to carbon biomass, the contribution of Rhizaria was highest between approximately 40° N and 20° S and was similar to the biomass of meso-zooplankton in the same latitudinal range⁵ (Fig. 2 and Extended Data Table 3). Overall, we estimate that the biomass of Rhizaria larger than 600 μm represents a standing stock of 0.089 Pg of carbon in the upper 200 m of the water column of the world ocean (Table 1). This biologically active carbon reservoir represents 29% of the combined meso- and macro-zooplankton biomass and 5.2% of the total oceanic biota carbon standing stock (Table 1). Despite the intrinsic uncertainties associated with such global assessments⁵, these values are consistent with previous estimates based on local studies²⁰.

The four main categories of Rhizaria discriminated in our analysis (Acantharia, Collodaria, Phaeodaria, and other Rhizaria; Extended Data Figs 2–4) exhibited distinct latitudinal biomass distributions (Extended Data Fig. 5). Overall, Phaeodaria and Collodaria were the most important contributors to rhizarian biomass, and Acantharia occurred at consistently low levels. However, most acantharian species are smaller than 600 μm and were therefore not quantified by our approach. The highest biomass of Collodaria and Acantharia occurred at low latitudes, whereas the biomass of Phaeodaria and other Rhizaria was more evenly distributed, suggesting that these orders of Rhizaria show distinct ecological preferences. In addition to latitudinal patterns, there was also a significant shift in taxonomic composition with depth (Fig. 3). In the top 100 m of the water column, photosymbiotic Collodaria contributed most to rhizarian biomass (Fig. 3a). Below, in the twilight zones of the oceans (depth 100–500 m), the asymbiotic phaeodarians were the most important contributors to rhizarian biomass at all latitudes (Fig. 3b). Considering that all Collodaria and most Acantharia investigated so far harbour symbiotic microalgae^{13,14,24}, we estimated that these groups typically contribute 0.18% (0.17% for Collodaria and 0.01% for Acantharia) of total primary production in oligotrophic waters (Extended Data Table 4). Only one study, performed in the oligotrophic Sargasso Sea, has estimated the contribution of Rhizaria to total primary production; this study showed that large photosymbiotic Rhizaria could account for 0.1–0.4% of total primary production²⁵. Even though the contribution of large photosymbiotic Rhizaria to total primary production is rather low, it occurs in the large size fraction, representing primary production directly available to large consumers and thereby shortcutting trophic levels of marine food webs¹⁴. Our sampling was restricted to organisms >600 μm and therefore excluded the abundant smaller species of rhizarians²⁴ and the top 5 m of the water column, where rhizarians can be highly abundant^{14,24,25}. Our estimates of rhizarian abundance and biomass should therefore be considered as conservative and further efforts are required to refine the emerging image of the global rhizarian contribution to biomass, primary productivity and other biogeochemical processes in the oceans.

The previously overlooked contribution of giant rhizarian biomass to plankton communities changes our perception of the oligotrophic tropical oceans. These oceans represent one of the largest ecosystems on the planet, occupying nearly 40% of the Earth's surface⁶, and are important biomes for the functioning of the biosphere. The use of appropriate tools provides new insights into global zooplankton community structure in the ocean, for instance, demonstrating that the

Table 1 | Carbon standing stock of giant Rhizaria in the 0–100, 0–200 and 0–500 m depth layers of the oceans

Depth layer (m)	Number of sampling stations	Rhizarian integrated biomass (mg C m ⁻²)				Global rhizarian biomass (Pg C)	Contribution to global:		
		Min	Max	Median	IQR		Carbon standing stock	Biomass of meso- and macro-zooplankton	Biomass of meso-zooplankton
0–100	877	0	23,910	34.43	247	0.012	-	-	-
0–200	848	0	146,400	245	1,219	0.089	5.2% (0.6–22%)	29% (4–68%)	31% (5–69%)
0–500	694	0	115,091	564	1,608	0.204	-	-	-

Estimates of global giant rhizarian carbon biomass were derived from median values assuming an ocean surface of $3.61 \times 10^{14} \text{ m}^2$. Giant rhizarian biomass contributions to global carbon and meso- and macro-zooplankton standing stocks were calculated on the basis of the median values for the 0–200 m depth layer (that is, only matching data available globally) published in ref. 5. The ranges of contribution were computed using the first and third quartiles of rhizarian integrated biomass. IQR, interquartile range. Global biomass estimates are expressed in petagrams of carbon (1 Pg = 10^{15} g). Detailed computational processes are provided in the Methods.

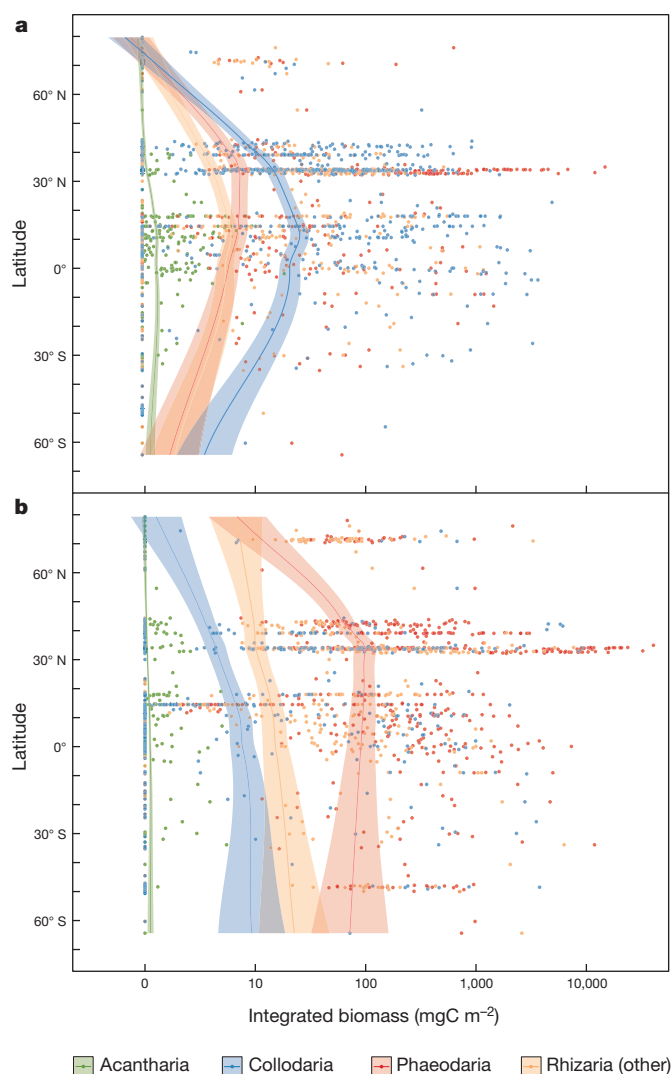


Figure 3 | Latitudinal distribution of depth integrated biomass ($\text{mg carbon (mg C) m}^{-2}$) for the different rhizarian taxa identified.
a, Biomass integrated in the top 100 m of the water column (877 sampling stations). **b**, Biomass integrated between 100 and 500 m depth (694 sampling stations). Latitudinal trends are represented by computing Loess regressions with a polynomial fitting. Shaded areas represent 95% confidence intervals.

abundance of photosymbiotic Rhizaria declines less markedly than that of other, non-photosymbiotic, zooplankton along a trophic gradient (Extended Data Fig. 6), and thus emphasizing the idea that photosymbiosis allows these large organisms to thrive in otherwise hostile oligotrophic environments^{11,26}. To date, large rhizarians have been omitted from biogeochemical flux budgets, but they may be efficient vectors for fluxes to the deep ocean via both primary production and vertical flux and could be important components of the biological pump¹⁵. Along with other cryptic, fragile and transparent creatures such as gelatinous plankton organisms, whose abundance probably remains poorly quantified²¹, rhizarians may thus contribute to carbon budgets in the dark mesopelagic ocean. The measured activity of microbial remineralization in the dark mesopelagic ocean exceeds the estimated carbon input, emphasizing the need to understand the synergy between microbes and large zooplankton to understand the processes that control the oceanic carbon sink^{27,28}. Along with better spatio-temporal descriptions of the occurrence of specific taxa (for example, Phaeodaria in the California coastal upwelling), accurate estimates of poorly known processes such as grazing, growth and biomineralization of Rhizaria and the photophysiology and carbon fixation of their symbionts are

required to allow us to include this significant component of the oceanic biota in ecological and biogeochemical models at both local and global scales.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 May 2015; accepted 10 March 2016.

Published online 20 April 2016.

- Behrenfeld, M. J. *et al.* Climate-driven trends in contemporary ocean productivity. *Nature* **444**, 752–755 (2006).
- Beaugrand, G., Edwards, M. & Legendre, L. Marine biodiversity, ecosystem functioning, and carbon cycles. *Proc. Natl Acad. Sci. USA* **107**, 10120–10124 (2010).
- Buitenhuis, E. *et al.* Biogeochemical fluxes through mesozooplankton. *Glob. Biogeochem. Cycles* **20**, GB2003 (2006).
- Rombouts, I. *et al.* Global latitudinal variations in marine copepod diversity and environmental factors. *Proc. R. Soc. Lond. B* **276**, 3053–3062 (2009).
- Buitenhuis, E. T. *et al.* MAREDAT: towards a world atlas of MARine Ecosystem DATA. *Earth Syst. Sci. Data* **5**, 227–239 (2013).
- Polovina, J. J., Howell, E. A. & Abecassis, M. Ocean's least productive waters are expanding. *Geophys. Res. Lett.* **35**, L03618 (2008).
- Banase, K. Zooplankton: Pivotal role in the control of ocean production. *ICES J. Mar. Sci.* **52**, 265–277 (1995).
- Wilson, S. E., Ruhl, H. A. & Smith, K. L. Zooplankton fecal pellet flux in the abyssal northeast Pacific: A 15 year time-series study. *Limnol. Oceanogr.* **58**, 881–892 (2013).
- Le Quéré, C. *et al.* Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. *Glob. Change Biol.* **11**, 2016–2040 (2005).
- Burki, F. & Keeling, P. J. Rhizaria. *Curr. Biol.* **24**, R103–R107 (2014).
- Stoecker, D. K., Johnson, M. D., de Vargas, C. & Not, F. Acquired phototrophy in aquatic protists. *Aquat. Microb. Ecol.* **57**, 279–310 (2009).
- De Wever, P., Dumitrica, P., Caulet, J. P., Nigrini, C. & Caridroit, M. *Radiolarians in the Sedimentary Record* (Taylor & Francis, 2001).
- Suzuki, N. & Not, F. in *Marine Protists* (eds Ohtsuka, S., Suzuki, T., Horiguchi, T., Suzuki, N. & Not, F.) 179–222 (Springer Japan, 2015).
- Anderson, O. R. *Radiolaria* (Springer, 1983).
- Lampitt, R. S., Salter, I. & Johns, D. Radiolaria: Major exporters of organic carbon to the deep ocean. *Glob. Biogeochem. Cycles* **23**, GB1010 (2009).
- Fontanez, K. M., Eppley, J. M., Samo, T. J., Karl, D. M. & DeLong, E. F. Microbial community structure and function on sinking particles in the North Pacific Subtropical Gyre. *Front. Microbiol.* **6**, 469 (2015).
- de Vargas, C. *et al.* Eukaryotic plankton diversity in the sunlit global ocean. *Science* **348**, 1261605 (2015).
- Lima-Mendez, G. *et al.* Top-down determinants of community structure in the global plankton interactome. *Science* **348**, 1262073 (2015).
- Guidi, L. *et al.* Plankton community and gene networks associated with carbon export in the global ocean. *Nature* (in the press).
- Dennett, M. R., Caron, D. A., Michaels, A. F., Gallagher, S. M. & Davis, C. S. Video plankton recorder reveals high abundances of colonial Radiolaria in surface waters of the central North Pacific. *J. Plankton Res.* **24**, 797–805 (2002).
- Remsen, A., Hopkins, T. L. & Samson, S. What you see is not what you catch: a comparison of concurrently collected net, Optical Plankton Counter, and Shadowed Image Particle Profiling Evaluation Recorder data from the northeast Gulf of Mexico. *Deep Sea Res. Part I Oceanogr. Res. Pap.* **51**, 129–151 (2004).
- Stemmann, L. *et al.* Global zoogeography of fragile macrozooplankton in the upper 100–1000 m inferred from the underwater video profiler. *ICES J. Mar. Sci.* **65**, 433–442 (2008).
- Picheral, M. *et al.* The Underwater Vision Profiler 5: An advanced instrument for high spatial resolution studies of particle size spectra and zooplankton. *Limnol. Oceanogr. Methods* **8**, 462–473 (2010).
- Michaels, A. F. Vertical distribution and abundance of Acantharia and their symbionts. *Mar. Biol.* **97**, 559–569 (1988).
- Caron, D. A., Michaels, A. F., Swanberg, N. R. & Howse, F. A. Primary productivity by symbiont-bearing planktonic saccodines (Acantharia, Radiolaria, Foraminifera) in surface waters near Bermuda. *J. Plankton Res.* **17**, 103–129 (1995).
- Taylor, F. J. R. in *The Ecology of Marine Protozoa* (ed. Capriulo, G. M.) 323–340 (Oxford Univ. Press, 1990).
- Herndl, G. J. & Reinthaler, T. Microbial control of the dark end of the biological pump. *Nature Geosci.* **6**, 718–724 (2013).
- Giering, S. L. C. *et al.* Reconciliation of the carbon budget in the ocean's twilight zone. *Nature* **507**, 480–483 (2014).
- Pesant, S. *et al.* Tara Oceans Data: A sampling strategy and methodology for the study of marine plankton in their environmental context. *Sci. Data* **2**, 150023 (2015).
- Longhurst, A. *Ecological Geography of the Sea* (Academic, 2010).
- Moriarty, R. & O'Brien, T. D. Distribution of mesozooplankton biomass in the global ocean. *Earth Syst. Sci. Data* **5**, 45–55 (2013).

Acknowledgements Thanks to J.-O. Irisson for help with the R language and statistical analysis and I. Probert and J. Dolan for comments and English proofreading. The following people were involved in cruise organization: T. Moutin (BOUM), M. Landry and M. Ohman (CCE LTER), S. Blain (KEOPS II), V. Smetacek and W. Naqvi (LOHAFEX), J. Karstensen (M96), M. Babin (Malina), L. Coppola (Moose GE), P. Brandt (MSM22) and M. Visbeck (MSM23). The following people were involved in plankton image sorting: L. Burdorf (CNRS LOV), C. Desnos (CNRS LOV), A. Forest (Tackuvit), G. IdAoud (CNRS LOV), M. P. Jouandet (MIO Pytheas), J. Poulain (CEA), J. Baptiste Romagnan (CNRS LOV), F. Roullier (CNRS LOV), S. Searson (CNRS LOV), B. Serranito (EBMA-PROTEE) and N. Vasset (CNRS LOV). This study is a contribution from the CCE-LTER program, supported by the U.S. National Science Foundation. For the *Tara* Oceans expedition we thank the CNRS (in particular Groupement de Recherche GDR3280), European Molecular Biology Laboratory (EMBL), Genoscope/CEA, VIB, Stazione Zoologica Anton Dohrn, UNIMIB, Fund for Scientific Research – Flanders, Rega Institute, KU Leuven and the French Ministry of Research. We also thank A. Bourgois and E. Bourgois, the Veolia Environment Foundation, Région Bretagne, Lorient Agglomération, World Courier, Illumina, the EDF Foundation, FRB, the Prince Albert II de Monaco Foundation, the *Tara* schooner and its captains and crew. We are also grateful to the French Ministry of Foreign Affairs for supporting the expedition and to the countries who granted sampling permission. *Tara* Oceans would not exist without continuous support from 23 institutes (<http://oceans.taraexpeditions.org>). The authors further declare that all data reported herein are fully and freely available from the date of publication,

with no restrictions, and that all of the samples, analyses, publications, and ownership of data are free from legal entanglement or restriction of any sort by the various nations in whose waters the *Tara* Oceans expedition sampled. Data described herein are available at PANGAEA (<http://doi.pangaea.de/10.1594/PANGAEA.842227>), and the data release policy regarding future public release of *Tara* Oceans data is described in ref. 29. Funding was from DESIR project Emergence-UPMC from Université Pierre et Marie Curie, JST-CNRS exchange program, CHAIRE CNRS/UPMC Vision, Investissements d'Avenir' programmes OCEANOMICS (ANR-11-BTBR-0008), DFG through SFB754 (GEOMAR and Kiel University) and Future Ocean (Kiel University and GEOMAR). This article is contribution number 38 from *Tara* Oceans.

Author Contributions F.N. and L.S. designed the study. M.P., T.B., R.K., P.V., H.H., N.M. and G.G. acquired and extracted raw data. T.B. produced the morphological classification of the rhizarian UVP images. T.B., L.S. and L.G. performed statistical analyses. R.K. and T.B. calculated the primary production contributions. F.N. and T.B. wrote the manuscript and produced display items. L.S., R.K., L.G., M.P., H.H. and G.G. discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.S. (stemmann@obs-vlfr.fr) or F.N. (not@sb-roscoff.fr).

METHODS

Sampling sites. Rhizarian distribution was observed with the Underwater Vision Profiler 5 (UVP5) deployed at 877 stations distributed across all oceans (11 cruises corresponding to 1,454 profiles). Out of these stations, 694 were sampled down to 500 m depth (Fig. 1 and Extended Data Table 1). Stations encompassed regions with a broad range of oceanographic structures (upwelling, boundary currents, large tropical gyres, etc.) from oligotrophic to eutrophic ecosystems. Sampling effort occurred throughout the year between 2008 and 2013 and covered latitudes from 65° S to 75° N. The majority of stations (72%) were sampled between 5° S and 40° N (Extended Data Fig. 1).

UVP5 deployments and raw data collection. The UVP5 images large plankton (equivalent spherical diameter, ESD > 600 µm; ref. 23). The UVP5 sampling volume varied from 0.5 to 1 l and images were recorded every 5 to 20 cm along vertical profiles, leading to an observed volume of 5 m³ for a 500 m depth profile. Mounted on a CTD rosette frame, the UVP5 starts recording below 5 m, ultimately leading to an underestimation in the quantification of objects just beneath the sea surface. Images produced by the UVP5 were extracted using the ZooProcess software³². For all objects, the major and minor axes of the best fitting ellipses were measured. A computer-assisted method was used to classify all organisms. Image identification was possible for objects larger than 600 µm. All images (total number ~1.8 million) were checked by experts to discriminate Rhizaria (~36,000 images) from other plankton and detritus. In the present dataset, the maximum ESD recorded was 7 cm (for a ctenophore); among Rhizaria only, the maximum ESD was 2.5 cm (for a colonial collodarian) (Extended Data Figure 4). Thereafter, Rhizaria were classified into finer taxonomic levels for all profiles included in this work.

Refining the rhizarian image categories. The entire UVP5 image collection was scanned to infer the diversity of images associated with the Rhizaria. Using taxonomic expertise, ten categories were created and affiliated to known rhizarian taxa (Extended Data Fig. 2). Differences in shapes and grey level were used to distinguish between categories. Phaeodaria were divided into three categories: PhaL, PhaSe and PhaSt. The PhaSe category (Extended Data Fig. 2a) are small grey spheres (<5 mm) with a tiny black nucleus inside. The black dot is usually found in the centre of the sphere, but its position can vary. The edges of the sphere are darker than the interior (this is an important differentiating criterion from solitary collodarians). PhaSe phaeodarians can be found in aggregates consisting of tens of specimens. The PhaSt category (Extended Data Fig. 2b) are dark grey spheres with a large black or white nucleus. The interior of the sphere is entirely grey, unlike other phaeodarian categories. Tiny spines surround the sphere. Images of the PhaL category (Extended Data Fig. 2c) are characterized by multiple long spine-like extensions originating from a dark centre. This dark centre can be a simple black dot or a grey sphere with a dark nucleus. Acantharians (Acn; Extended Data Fig. 2e) possess short- to medium-sized spines surrounding a black centre. The most important criterion to distinguish acantharians from phaeodarians in the PhaL category in the UVP5 images is the symmetry of the spines, which is characteristic of acantharian cells. Collodarians were divided into five categories, including colonial and solitary specimens. Large collodarian colonies (Col; Extended Data Fig. 2j) are easily recognizable by their large size (often over 3 mm; Extended Data Fig. 4) and pigmented appearance. The colony shape is variable and can be spherical, stretched, an assemblage of spheres, and so on. A pale halo often surrounds the colony and is helpful for identifying collodarian colonies. The SolGlob category of solitary collodarians (Extended Data Fig. 2i) are large, spherical-to-oval organisms with a homogenous grey (or dark-grey) surface. The central sphere is surrounded by a blurry halo generated by a network of pseudopodial extensions. The SolB and SolG categories of solitary collodarians (Extended Data Fig. 2f, h) are large spherical organisms with a dense central part (black/dark grey and grey, respectively). A gradient of grey is observed from the central part to the outer part of the cell. As the outermost part of this halo is almost transparent, the edges of the organism are not always visible. The grey level of the central part is the main distinguishing criterion between the two categories. The last category of solitary collodarians (SolF; Extended Data Fig. 2g) also comprises large spherical organisms with a grey central part surrounded by a dark-grey fuzzy structure. A gradient of grey is observed from the central part to the outer part of the cell. The fuzzy structure around the central part is the main criterion that distinguishes these organisms from other solitary categories. Not all rhizarian images fitted into the categories defined above. The category Rhiz (Extended Data Fig. 2d) comprised rhizarians that could not be precisely fitted into the previous categories, such as Foraminifera, which also belong to the super-group Rhizaria but mostly fall below the size threshold of our camera system.

Qualitative calibration of the newly defined categories was performed on plankton samples collected gently in Villefranche-sur-Mer bay (France, 43°41'10" N, 7°19'00" E) using a Regent net (680 µm mesh size) and off California (Californian Current Ecosystem) using a 333 µm mesh size plankton net hauled at a maximum

speed of 0.5 m s⁻¹ to minimize damage to the specimens. Live rhizarian specimens (Collodaria and Phaeodaria) were handpicked from the samples and then transferred into 0.2 µm filtered seawater. Each specimen was identified and photographed. The UVP5 was immersed in an aquarium filled with 0.2 µm filtered seawater. Freshly isolated specimens were dropped one by one on top of the illuminated volume of water to capture *in situ* images. Comparison between *ex situ* and *in situ* images confirmed the categories defined for the UVP5 image collection (Extended Data Fig. 3).

Data analysis. Analyses of rhizarian data included five steps: (i) vertical binning of each profile into four depth layers (0–100 m, 0–200 m, 100–500 m and 0–500 m). These layers were selected on the basis of photic properties and availability of published matching plankton datasets for comparison and to maximize the number of sampling stations considered in our dataset. Then, for each depth layer, we calculated (ii) the integrated abundance and (iii) biomass of all Rhizaria, and (iv) the primary production by photosymbiotic Rhizaria. (v) The results were averaged according to the different biogeochemical regions. All analyses were performed in R (ref. 33) with the package ggplot2 (ref. 34).

(i) For each sampling station with several vertical profiles, all profiles were summed to construct one single profile. Stations were divided into two categories according to maximum deployment depth. The UVP5 recorded images down to 100 m depth in 877 stations, and down to 500 m depth in 694 stations (Extended Data Table 1). Sample volume was on average 1.74 ± 0.59 m³ between 0 and 100 m depth and 2.95 ± 0.81 m³ between 100 and 500 m depth.

(ii) Mean integrated abundances and relative contributions of Rhizaria were calculated for the 0–500 m layer. To assess the contribution of Rhizaria to the entire zooplankton community, the abundance of other planktonic groups identified during the image process was computed. All other zooplankton imaged by the UVP5 were distinguished from non-living particles by semi-automatic annotation validated by experts; these organisms included copepods, crustaceans (shrimp-like, amphipod, cladoceran), gelatinous zooplankton (jellyfishes, ctenophores, siphonophores, salps), chaetognaths, appendicularia, molluscs, annelids and fish larvae. Other particles (detritus, aggregates, etc.) and phytoplankton (large diatoms, *Trichodesmium*, etc.) were removed from the computation.

(iii) Biomass estimations for the different rhizarian categories were inferred from organism measurements (major and minor axes of the best fitting ellipse) generated during ZooProcess image processing. These axes were used for biomass calculation instead of the Equivalent Spherical Diameter (ESD), as the use of the latter leads to an overestimation of biomass for large and elongated organisms (such as long colonial collodarians). Biovolume was first calculated from geometric shapes for all categories except colonial collodarians: spheres for Acantharia, prolate ellipsoid for all other categories (Extended Data Table 3). Areas (A_e) of a prolate ellipsoid were determined for colonial collodarians as follows:

$$A_e = 2\pi \left(\frac{\text{minor}}{2} \right)^2 + \frac{2\pi \left(\frac{\text{Major}}{2} \times \frac{\text{minor}}{2} \right)}{e} \arcsin(e) \quad (1)$$

where e is the eccentricity of an ellipse:

$$e = \frac{\sqrt{\left(\frac{\text{Major}}{2} \right)^2 - \left(\frac{\text{minor}}{2} \right)^2}}{\left(\frac{\text{Major}}{2} \right)} \quad (2)$$

Biovolumes and surface areas were then converted to biomass using carbon conversion factors from the literature^{20,35,36} (Extended Data Table 3).

(iv) Primary production of photosymbiotic rhizarians was estimated individually for all Acantharia and Collodaria observed between 0 and 500 m depth. While all collodarian species investigated have been described as photosymbiotic^{13,14}, the vast majority of large acantharian specimens found in the upper water column are known to harbour symbionts^{24,25}. Individual primary production (iPP) was estimated for each photosymbiotic rhizarian as a function of the biovolume²⁵:

$$\log(\text{iPP}) = 0.62 \times \log(\text{Biov}) - 4.33 \quad (3)$$

where Biov is the biovolume of the holobiont estimated from a prolate ellipsoid.

Assuming a reference temperature (T_{ref}) of 23.5°C (ref. 25), we applied a Q_{10} temperature coefficient of 1.88 (ref. 37) to correct for temperature effects on photosynthetic production and yield temperature-corrected individual primary production (iPP_T):

$$\text{iPP}_T = \text{iPP} \left[1.88^{\left(\frac{T_{\text{ref}} - T_{\text{ctd}}}{10} \right)} \right]^{-1} \quad (4)$$

where T_{ctd} is the *in situ* temperature measured by the CTD at the depth of each organism.

The 490 nm light attenuation coefficient $K_d(490)$ (m^{-1}) and photosynthetic active radiation (PAR; mol photons per m^2 per day) were used to estimate the available instantaneous radiation at depth. Satellite-derived average daily PAR, net primary production (NPP), chlorophyll *a* (Chl_a) and $K_d(490)$ (8-day averages at 4-km resolution) were downloaded from the Oregon University database (<http://www.science.oregonstate.edu/ocean.productivity/>). Average values for each station were calculated for the position occupied $\pm 0.1^\circ$ if the occupation date fell within the 8-day window of the satellite observation. The PAR attenuation coefficient $K_d(\text{PAR})$ was calculated according to Morel³⁸ as:

$$K_d(\text{PAR}) = 0.0864 + 0.884 K_d(490) - 0.00137[K_d(490)]^{-1} \quad (5)$$

The average instantaneous radiation available at the surface (iPAR_s in μmol photons per m^2 per s) during daytime was calculated as:

$$\text{iPAR}_s = \left(\frac{\text{PAR}}{\text{daylength}} \right) / 3,600 \times 10^6 \quad (6)$$

with daylength in h per day. From this data, the instantaneous radiation available at a specific depth z (iPAR_d) was calculated as:

$$\text{iPAR}_d = e^{(K_d(\text{PAR})z)} \text{iPAR}_s \quad (7)$$

The primary productivity measurements for radiolarians and acantharians used in equation (3) were conducted at surface light conditions near Bermuda²⁵. These conditions are likely to be saturating light conditions for symbiotic rhizarians³⁹. Very little photophysiological information is available for photosymbiotic Rhizaria. To calculate the decrease in primary productivity with decreasing light availability, the fraction of primary productivity possible at a given iPAR (fPP_{iPAR}) was calculated from the light saturation intensity ($I_k = 165 \mu\text{mol}$ photons per m^2 per s) observed for *Globigerinoides sacculifer*, a photosymbiotic planktonic Foraminifera³⁹ according to the hyperbolic tangent function for the light dependency of photosynthesis in marine phytoplankton⁴⁰ as:

$$\text{fPP}_{\text{iPAR}} = \tanh \left(\frac{\text{iPAR}_d}{I_k} \right) \quad (8)$$

The individual primary productivity at a given depth was then calculated as:

$$\text{iPP} = \text{iPP}_T \times \text{fPP}_{\text{iPAR}} \quad (9)$$

Finally, given the paucity of information available, it should be noted that we used size–primary productivity relations for acantharians and colpodarians from just one study²⁵ and the dependency of photosynthesis on light availability from a planktonic Foraminifera for these calculations.

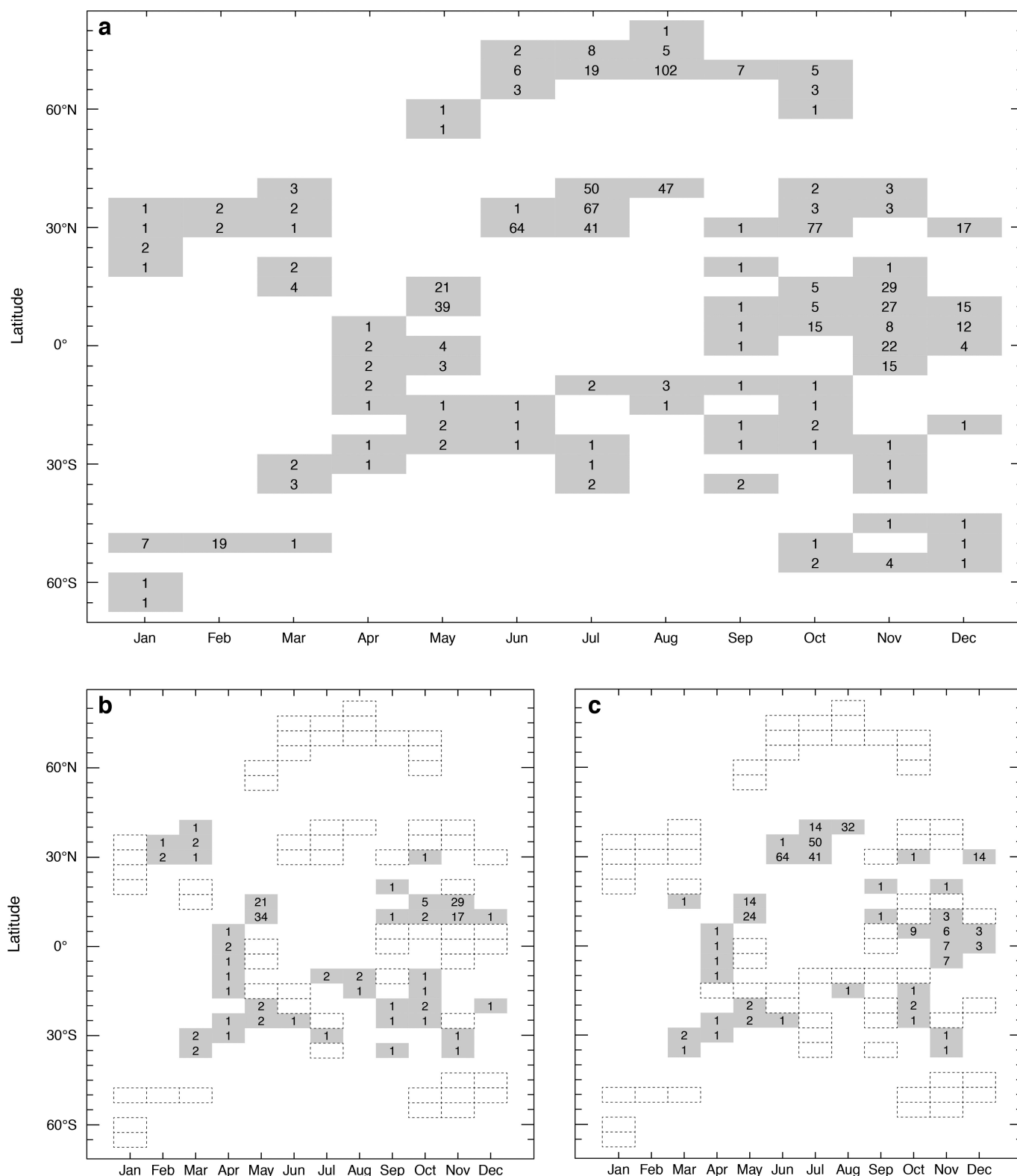
(v) Each UVP5 station was finally affiliated to one of 33 Longhurst's biogeochemical provinces and assembled in the biomes defined by Longhurst³⁰ (Extended Data Tables 1 and 2). Only 29 stations were sampled within Longhurst's coastal provinces (FKLD, BRAZ, CARM, CHIL, GUIA, NWCS, ARAB, REDS and EAFR). The bottom depth at these stations was always more than 500 m and they were not located on the continental shelf or continental slope. All these coastal province stations were therefore merged with their adjacent oceanic biomes, and the Antarctic Biome included both the Antarctic Polar Biome and the Antarctic Westerly Winds Biome. The merging of sampling stations located in Longhurst's coastal provinces with their adjacent biomes did not affect the contribution of Rhizaria to zooplankton communities in these biomes. Indeed, when considered separately, the average Rhizaria contribution for coastal provinces reached similar values as for the global dataset (33.81%). Two provinces, the Mediterranean Sea (MEDI) and the California Upwelling Coastal (CCAL) provinces, were treated separately from their respective biomes (Atlantic Coastal and Pacific Coastal biomes) because both were densely sampled and showed high rhizarian abundances compared to the other provinces in the same biomes.

Global estimates. Global estimates of rhizarian biomass were computed for three different layers (Table 1) assuming an ocean surface of $3.61 \times 10^{14} \text{ m}^2$. All estimates were derived from median biomass values to prevent overestimates using mean values, the latter being highly influenced by locally high biomass values (for example, the California Current). Low and high estimates of global rhizarian biomass were

computed using the first and third quartile, respectively. We compared the global rhizarian biomass to independent data on the global average estimates of meso- and macro-zooplankton biomass in the first 200 m of the oceans⁵. The contribution of Rhizaria to global plankton biomass was established using 11 different plankton functional types (PFTs)⁵, including autotrophic and heterotrophic PFTs. The median derived biomass for each plankton group was considered in the top 200 m and summed to provide an estimate of the plankton carbon standing stock. Despite uncertainties inherent to any global estimates (for example, carbon conversion factors, sampling coverage; ref. 5) we intended to provide a conservative contribution of Rhizaria to the different plankton components. The relative contribution of global rhizarian biomass to global plankton carbon standing stock was therefore calculated as the global rhizarian biomass divided by the sum of the published reference estimate for global plankton biomass⁵ and the global rhizarian biomass. **Possible impact of sampling coverage on rhizarian biomass distribution pattern.** The global patterns observed in this study are inevitably associated with the sampling effort and geographic coverage (Extended Data Fig. 1). Some oceanic areas and/or seasons were more intensely sampled than others, creating heterogeneity in the dataset. For instance, the Mediterranean Sea and the California Current were sampled intensely. Although we sampled 33 of 51 Longhurst's provinces, our spatial coverage was partial. To assess the possible influence of sampling coverage on latitudinal pattern of rhizarian biomass distribution, we used the *sample* function (implemented in R version 3.2.0) to obtain a random subset of our dataset and tested the latitudinal pattern of this dataset against the original dataset. We selected five latitude intervals of 30° (between 90° N and 60° S) and randomly extracted 20 sampling stations from each with bootstrap resampling. The difference between the resampled dataset ($n = 100$ sampling stations) and the original entire dataset ($n = 694$ sampling stations) was tested with a non-parametric Mann–Whitney *U*-test and no significant difference was observed ($P = 0.155$).

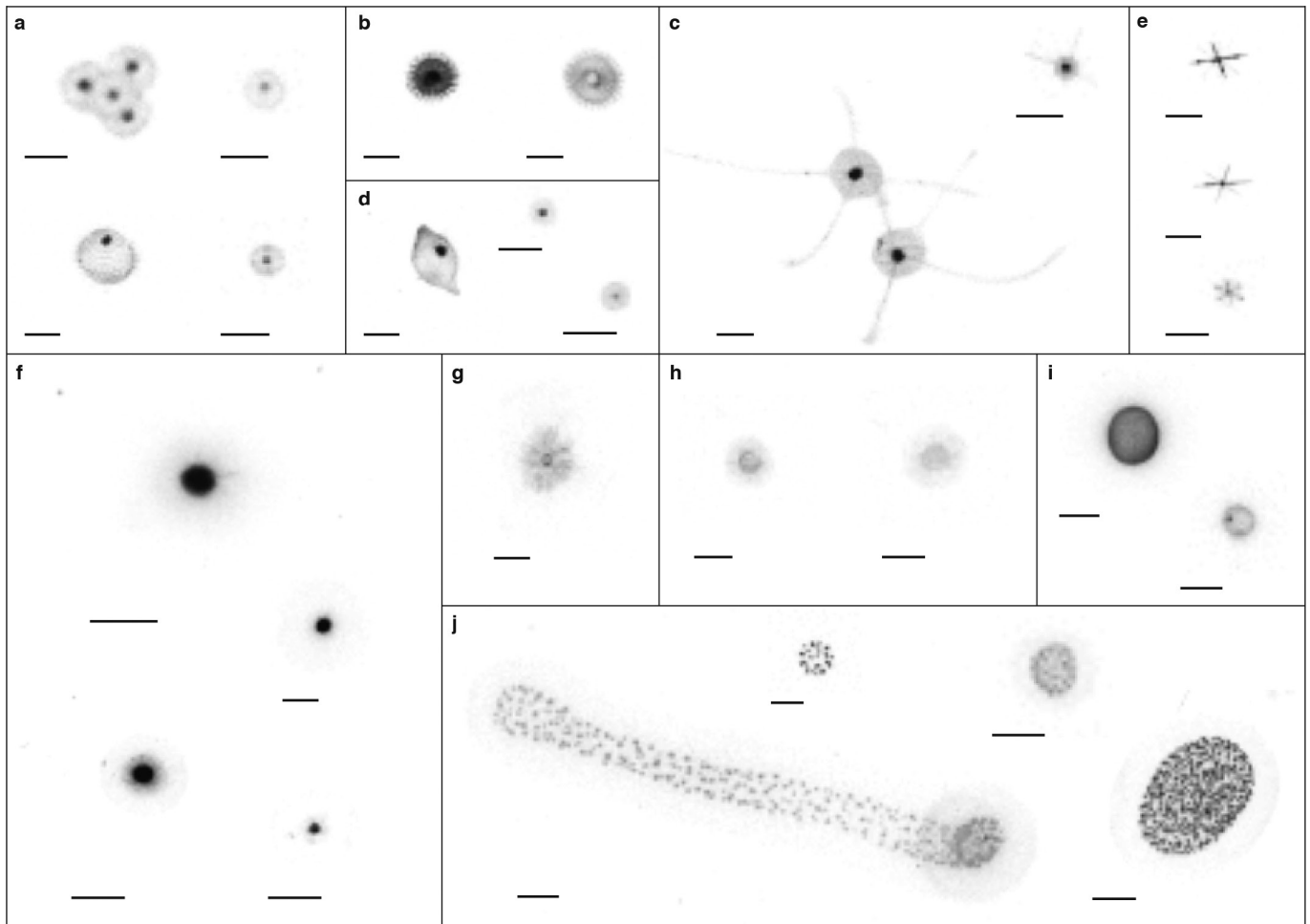
Data accessibility. The data described herein are publicly available at PANGAEA^{41–43}.

32. Gorsky, G. *et al.* Digital zooplankton image analysis using the ZooScan integrated system. *J. Plankton Res.* **32**, 285–303 (2010).
33. R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing* (2014).
34. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer Science & Business Media, 2009).
35. Michaels, A. F., Caron, D. A., Swanberg, N. R., Howse, F. A. & Michaels, C. M. Planktonic sarcodines (Acantharia, Radiolaria, Foraminifera) in surface waters near Bermuda: abundance, biomass and vertical flux. *J. Plankton Res.* **17**, 131–163 (1995).
36. Beers, J. R. & Stewart, G. L. in *The Ecology of the Plankton off La Jolla, California, in the Period April through September, 1967, Part VI* (eds Strickland, J. D. H., Solarzano, L. & Eppley, R. W.) **17**, 67–87 (Bull. Scripps Inst. Oceanogr., 1970).
37. Bissinger, J. E., Montagnes, D. J. S., Sharples, J. & Atkinson, D. Predicting marine phytoplankton maximum growth rates from temperature: Improving on the Eppley curve using quantile regression. *Limnol. Oceanogr.* **53**, 487–493 (2008).
38. Morel, A. Examining the consistency of products derived from various ocean color sensors in open ocean (Case 1) waters in the perspective of a multi-sensor approach. *Remote Sens. Environ.* **111**, 69–88 (2007).
39. Jørgensen, B. B., Erez, J., Revsbech, N. P. & Cohen, Y. Symbiotic photosynthesis in a planktonic foraminiferan, *Globigerinoides sacculifer* (Brady), studied with microelectrodes. *Limnol. Oceanogr.* **30**, 1253–1267 (1985).
40. Jassby, A. D. & Platt, T. Mathematical formulation of the relationship between photosynthesis and light for phytoplankton. *Limnol. Oceanogr.* **21**, 540–547 (1976).
41. Biard, T. *et al.* Abundance of large protists from the Infrakingdom Rhizaria in the global ocean. PANGAEA. <http://dx.doi.org/10.1594/PANGAEA.858136> (2016).
42. Biard, T. *et al.* Biomass of large protists from the Infrakingdom Rhizaria in the global ocean. PANGAEA <http://dx.doi.org/10.1594/PANGAEA.858156> (2016).
43. Biard, T. *et al.* Environmental context of a compilation about the distribution of large protists from the Infrakingdom Rhizaria in the global ocean. PANGAEA <http://dx.doi.org/10.1594/PANGAEA.858158> (2016).
44. Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science* **281**, 237–240 (1998).
45. Richardson, T. L. & Jackson, G. A. Small phytoplankton and carbon export from the surface ocean. *Science* **315**, 838–840 (2007).



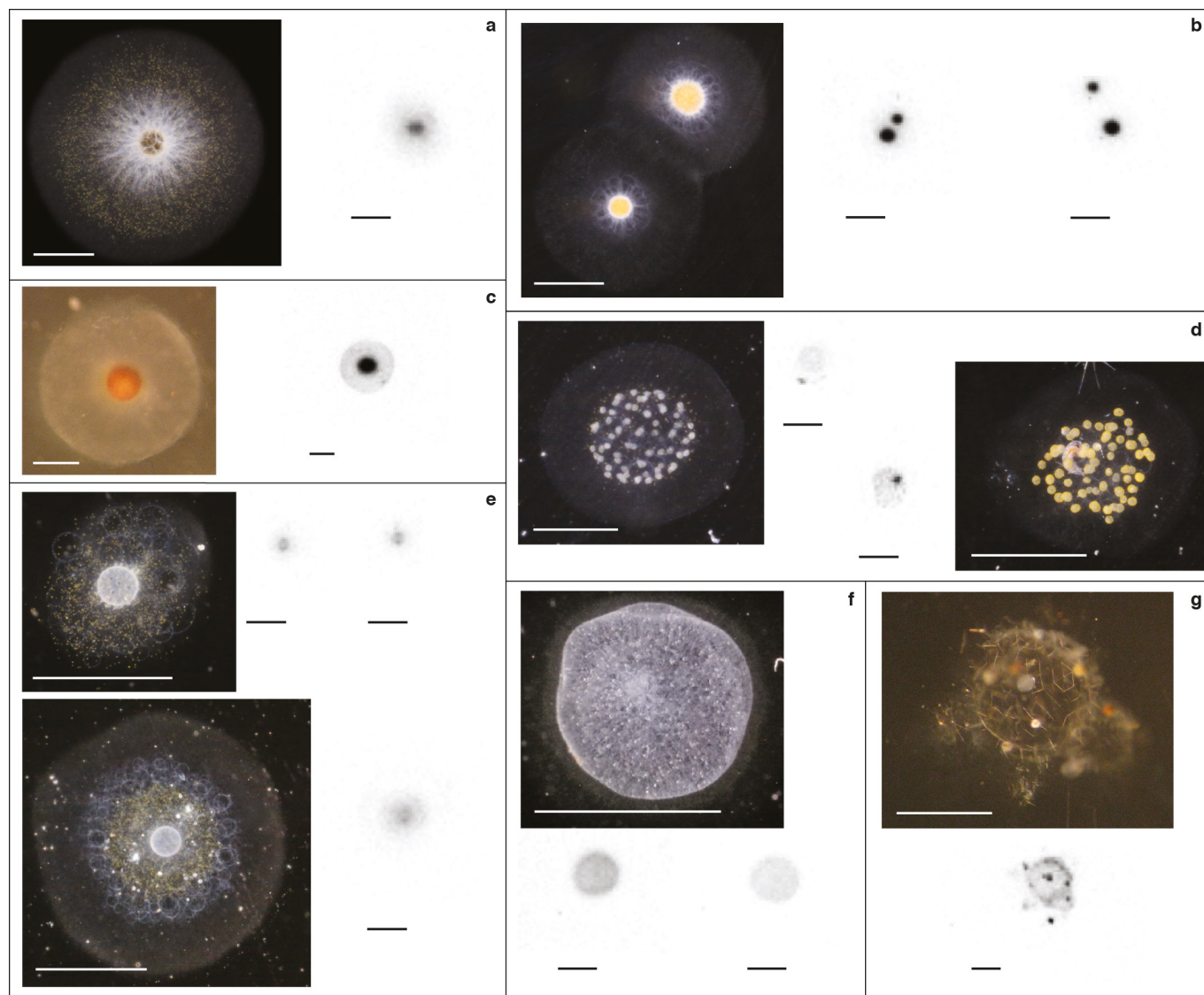
Extended Data Figure 1 | Sampling effort of the Underwater Vision Profiler surveys used in our study, represented across latitudes and months of the year. Rectangles identify latitude intervals of 5° affiliated to a given month. Numbers inside rectangles indicate the number of stations sampled. **a**, Sampling effort for the full dataset. **b**, Sampling

stations identified as belonging to one of Longhurst's gyral biogeochemical provinces. **c**, Sampling stations identified as belonging to oligotrophic waters ($\text{Chl}_a^{\text{sat}} < 0.1 \text{ mg m}^{-3}$; ref. 44). White rectangles with dashed edges highlight sampling stations not belonging to a gyre nor oligotrophic waters.



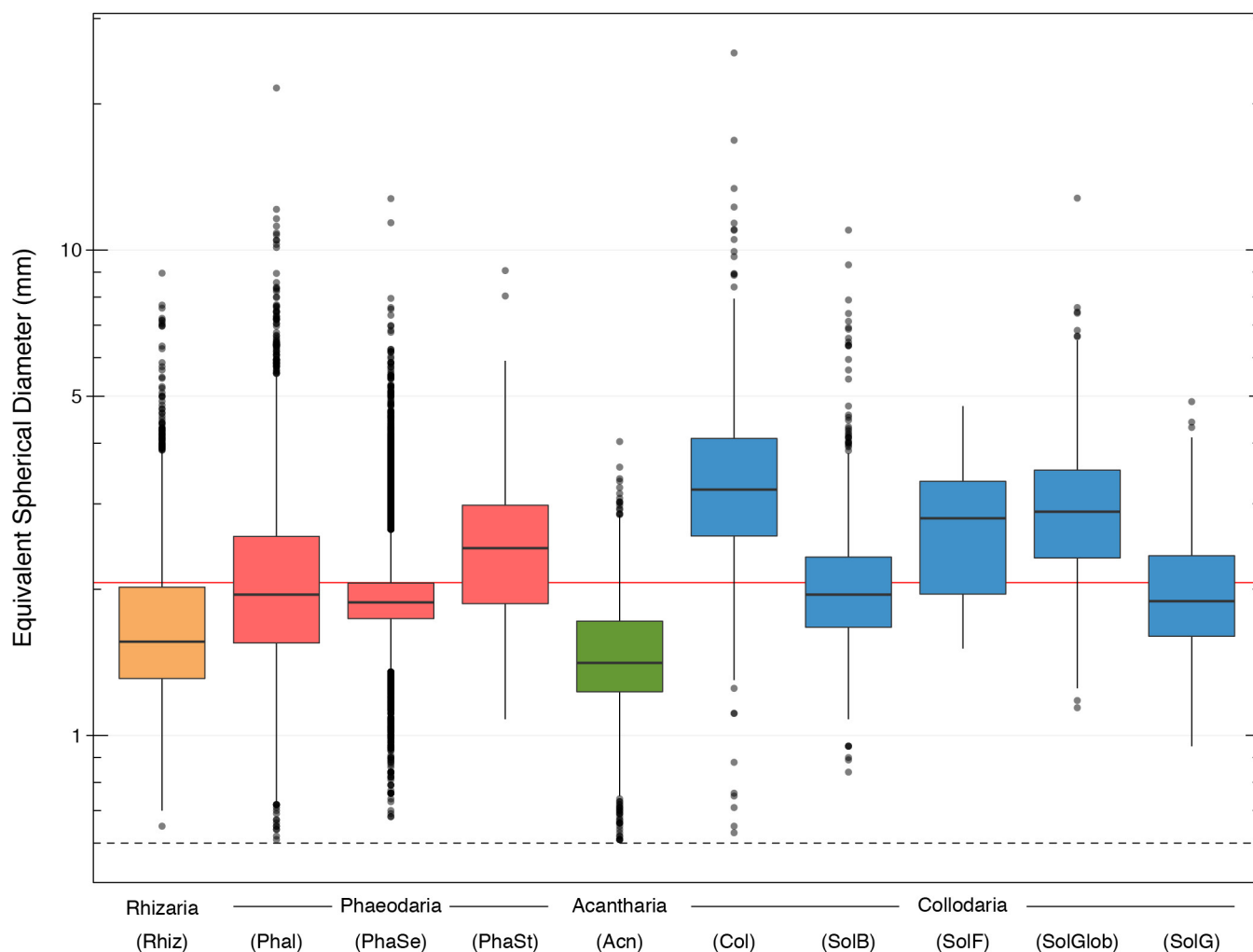
Extended Data Figure 2 | Images of the different rhizarian categories obtained with the UVP5. **a–c**, Phaeodaria: **(a)** phaeodarian spheres (PhaSe), **(b)** phaeodarian spheres with thorn edges (PhaSt) and **(c)** phaeodarians with long extensions (PhaL). **d**, Unidentified rhizarians (Rhiz). **e**, Acantharia (Acn). **f–j**, Collodaria: **(f)** solitary collodarians

with a dark central capsule (SolB), **(g)** solitary collodarians with a fuzzy central capsule (SolF), **(h)** solitary collodarians with a grey central capsule (SolG), **(i)** solitary collodarians with a globule-like appearance (SolGlob) and **(j)** colonial collodarians (Col). Detailed descriptions of the different categories are provided in the Methods. Scale bars, 2 mm.



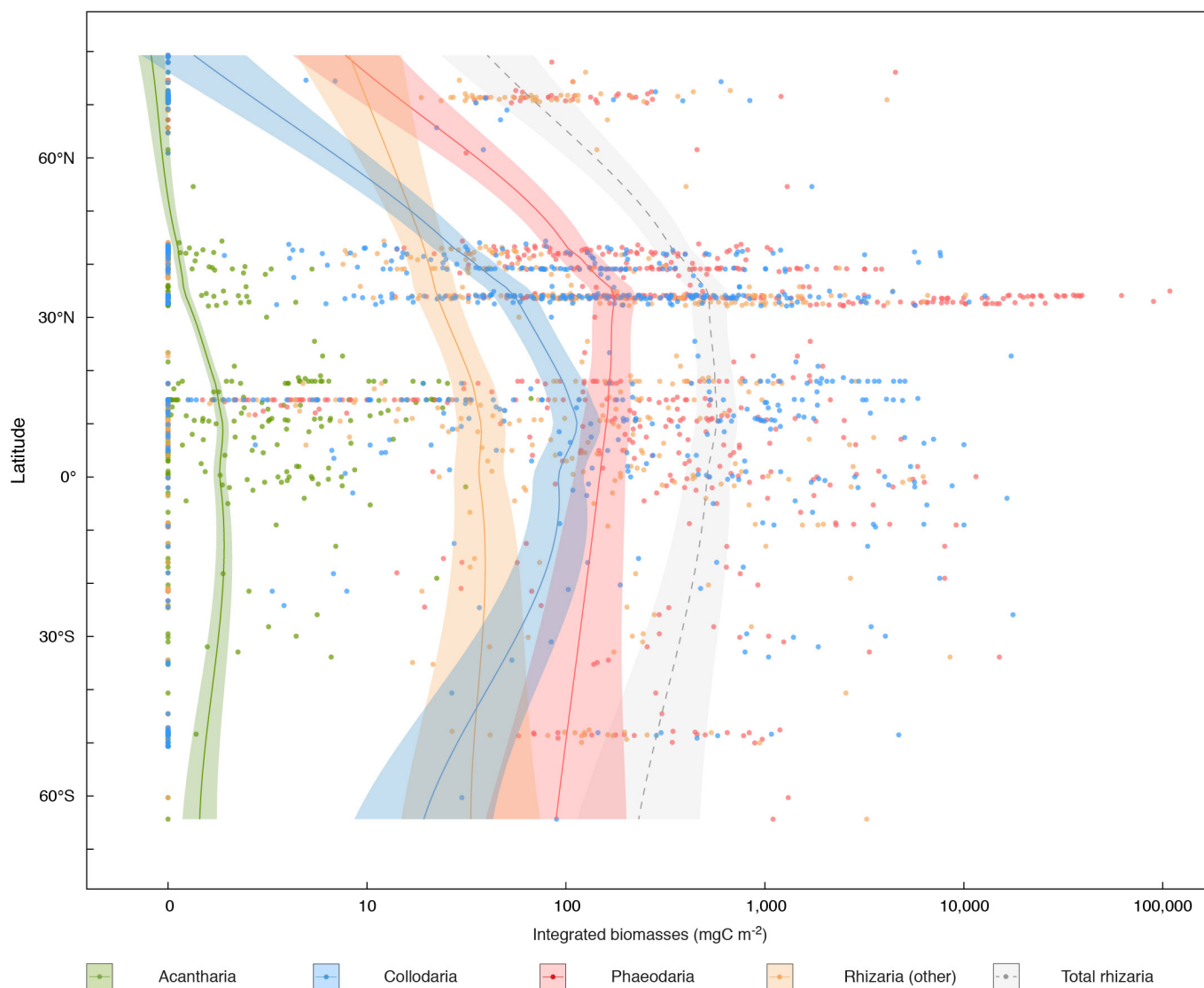
Extended Data Figure 3 | Calibration of rhizarian categories through comparison of single specimen images acquired by UVP5 and optical microscopy. Optical microscopy images and UVP5 images were obtained from the same specimens. **a**, *Thalassicolla caerulea* (SolB). **b**, **c**, Unidentified solitary collodarian species with dark central capsules

(SolB). **d**, Small collodarian colonies (Col). **e**, *Procyttarium primordialis* (two solitary collodarians with a white central capsule; SolG). **f**, *Physematium muelleri* (a solitary collodarian with a granular and opaque surface, similar to SolG). **g**, The Phaeosphaeridae family of Phaeodaria (PhaSe). Scale bars, 2 mm.



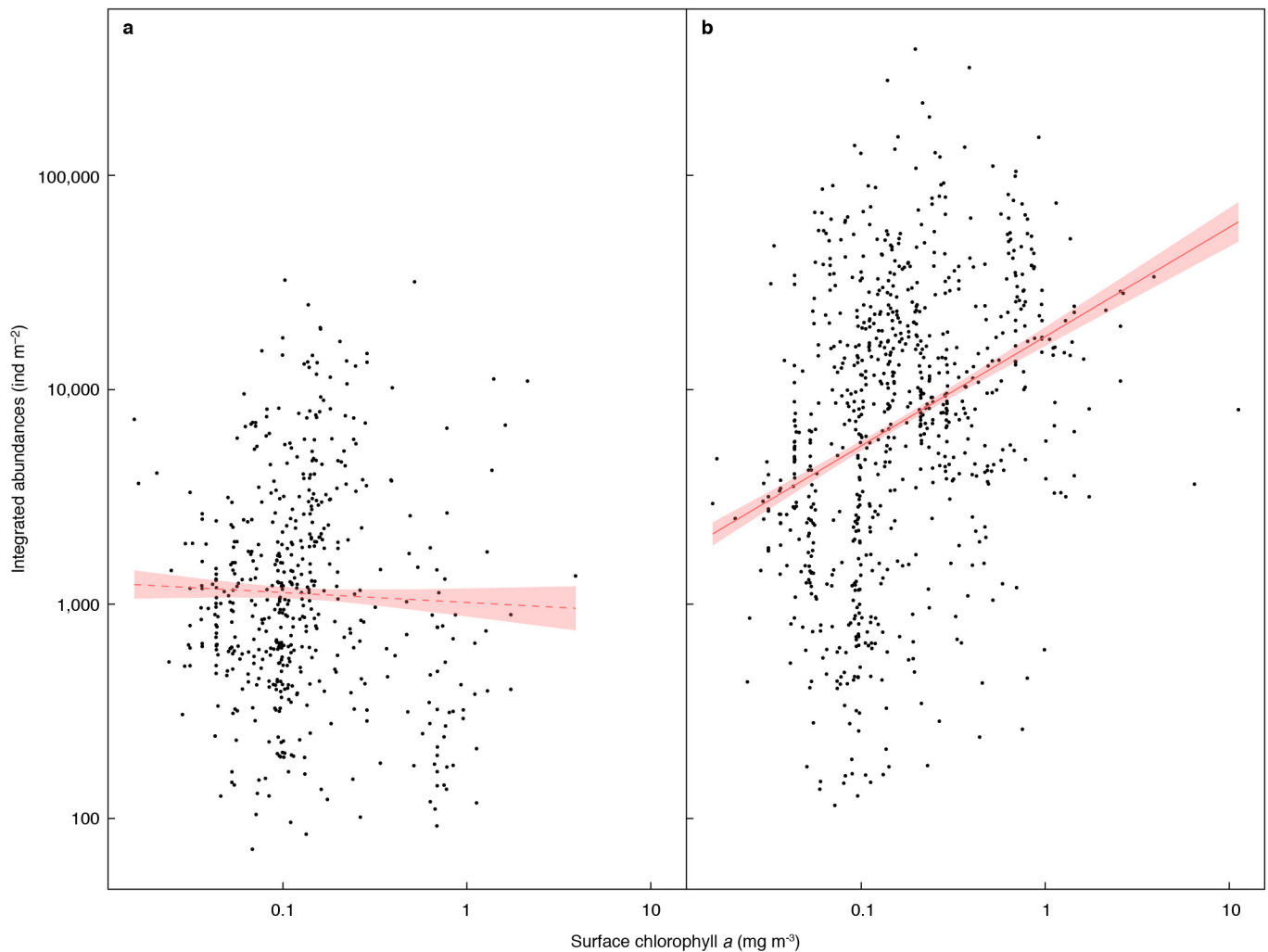
Extended Data Figure 4 | Size distribution of rhizarian categories in the UVP5 dataset. The dashed line represents the 600- μm size threshold of the camera. The overall mean equivalent spherical diameter (ESD) is 2.06 mm (red line). Dark horizontal lines represent the mean, boxes

represent the first and third quartiles for data distribution around the mean and the whiskers denote the lowest and highest values within 1.5 IQR from the first and third quartiles. Outlier values are represented by dots.



Extended Data Figure 5 | Latitudinal biomass distribution (mg C m^{-2}) of the different rhizarian taxa identified (Acantharia, Collodaria, Phaeodaria and other Rhizaria) integrated over the top 500 m of the

oceans (694 sampling stations). Loess regressions with polynomial fitting were computed to illustrate the latitudinal trends. Shaded areas represent 95% confidence intervals.



Extended Data Figure 6 | Variation in UVP5 depth-integrated abundances (0–100 m depth) as a function of the MODIS surface chlorophyll *a* extracted from satellite data (Oregon University Database). Solid and dashed red lines indicate significant and non-significant linear regressions, respectively. The shaded areas represent the standard error. **a**, The integrated abundance of photosymbiotic Rhizaria ($n = 521$) was not significantly linearly

dependent on chlorophyll *a* concentrations ($F = 0.622$, $R^2_{\text{adj}} = -0.0007$, $P = 0.431$). We assume that all collodarian species are photosymbiotic^{13,14} and that the majority of large acantharian cells found in the photic layer are known to harbour symbionts^{24,25}. **b**, The integrated abundance of other zooplankton (including asymbiotic Rhizaria; $n = 793$) decreased linearly along a trophic gradient ($F = 94.51$, $R^2_{\text{adj}} = 0.106$, $P < 10^{-16}$).

Extended Data Table 1 | Sampling cruise information, number of stations sampled, and number of UVP5 deployments (for example, profiles) used to generate the dataset analysed in this study

Sampling Cruise Name	Year	Chief Scientists	Biomes	(0-100 m)		(100-500 m)	
				Sampling Stations	Profiles	Sampling Stations	Profiles
BOUM	2008	T. Moutin	Mediterranean Sea* (MEDI)	183	183	151	151
KEOPS II	2011	S. Blain	Antarctic Biome (ANTA)	7	7	7	7
LOHAFEX	2009	V. Smetacek	Antarctic Biome (ANTA)	27	27	21	21
CCE-LTER	2008	M. Landry M. Ohman	California Upwelling Coastal* (CCAL)	75	75	58	58
M96	2013	J. Karstensen	Atlantic Trade Wind Biome (SATL)	60	77	58	59
MALINA	2009	M. Babin	Atlantic Polar Biome (ATPL)	119	119	54	54
MOOSE- GE	2012	L. Coppola	Mediterranean Sea* (MEDI)	87	87	79	79
MSM22	2012	P. Brandt	Atlantic Trade Wind Biome (SATL)	108	108	80	80
MSM23	2012	M. Visbeck	Atlantic Trade Wind Biome (SATL)	45	45	45	45
Tara Oceans	2009	Tara Oceans Consortium	Mediterranean Sea* (MEDI)	27	46	24	29
	2010		Indian Ocean Trade Wind Biome (IND)	23	94	22	80
	2010		Atlantic Trade Wind Biome (SATL)	16	82	14	55
	2011		Antarctic Biome (ANTA)	3	6	3	5
	2011		Pacific Trade Wind Biome (PAC)	34	218	33	140
	2011		California Upwelling Coastal* (CCAL)	4	30	4	26
	2012		Atlantic Trade Wind Biome (SATL)	2	17	2	13
	2012		Atlantic Westerly Winds Biome (NATL)	12	92	12	73
	2013		Atlantic Westerly Winds Biome (NATL)	1	9	1	9
	2013		Atlantic Polar Biome (ATPL)	44	132	26	72
Total				877	1,454	694	1,056

Biogeochemical biomes are defined according to ref. 30.

*This province was treated separately from its biome because it showed a strong pattern of rhizarian abundance compared to the other provinces in the same biome.

Extended Data Table 2 | Respective contributions of Rhizaria and other zooplankton abundances to the zooplankton community (>600 μm) integrated for the top 0–500 m of the water column**a**

Longhurst's Biogeochemical Biomes	Number of sampling stations	Number of profiles	Zooplankton community / Sampling station (ind m^{-2})	Contribution to zooplankton community (%)	
				Rhizaria	Other zooplankton
Atlantic Polar Biome (ATLP)	80	126	206,905	3	97
Antarctic Biome (ANTA)	31	33	126,802	4	96
Atlantic Westerly Winds Biome (NATL)	13	82	434,501	15	85
Atlantic Trade Wind Biome (SATL)	199	252	273,721	15	85
Pacific Trade Wind Biome (PAC)	33	140	354,035	35	65
Indian Ocean Trade Wind Biome (IND)	22	80	68,911	37	63
Mediterranean Sea* (MEDI)	254	259	21,269	47	53
California Upwelling Coastal* (CCAL)	62	84	556,582	81	19
Average proportion				33	67

b

Sampling station category	Number of sampling stations	Contribution to zooplankton community (%)	
		Rhizaria	Other zooplankton
'Gyre'	144	17	83
'Non-gyre'	550	17	83
Oligotrophic area	273	21	79
Non-oligotrophic area	408	17	83

a, Contributions in the different Longhurst's biogeochemical biomes. **b**, Contributions computed by removing sampling stations from the California coastal upwelling, for gyre/non-gyre and oligotrophic/non-oligotrophic ($\text{Chla} < 0.1 \text{ mg m}^{-3}$; ref. 44) categories.

*This province was treated separately from its biome because it showed a strong pattern of rhizarian abundance compared to the other provinces in the same biome.

Extended Data Table 3 | Carbon conversion factors used to assess biomass for the rhizarian categories discriminated

Category	Parameter estimated	Carbon conversion factors*	Conversion factors references
Acantharia (Acn)	Biovolume	0.0026 mgC mm ⁻³	(40)
Collodaria_colony (Col)	Surface Area	133 ngC cc ⁻¹	(20,40)
Collodaria_solitary_black (SolB)	Biovolume	0.28 mgC mm ⁻³	(40)
Collodaria_solitary_fuzzy (SolF)	Biovolume	0.28 mgC mm ⁻³	(40)
Collodaria_solitary_globule (SolGlob)	Biovolume	0.009 mgC mm ⁻³	(40)
Collodaria_solitary_grey (SolG)	Biovolume	0.28 mgC mm ⁻³	(40)
Phaeodaria_leg (PhaL)	Biovolume	0.08 mgC mm ⁻³	(41)
Phaeodaria_sphere_eye (PhaSe)	Biovolume	0.08 mgC mm ⁻³	(41)
Phaeodaria_sphere_thorn (PhaSt)	Biovolume	0.08 mgC mm ⁻³	(41)
Rhizaria_other (Rhiz)	Biovolume	0.08 mgC mm ⁻³	(41)

*Carbon contents are expressed as a function of the biovolume (mg C mm⁻³) or as a function of the number of central capsules (cc) in colonial collodarians.

Extended Data Table 4 | Net primary production of photosymbiotic giant rhizarians (Collodaria and Acantharia) and their contribution to total and >2- μ m net primary production in the global ocean and oligotrophic regions

	NPP per surface (mgC d ⁻¹ m ⁻²)			Contribution to global NPP (%)	Contribution in the oligotrophic regions (%)	
	Min	Max	Mean (\pm SEM)		To total NPP	To the >2 μ m size fraction
Photosymbiotic Rhizaria (0-500 m)	0	5.64	0.26 (\pm 0.03)	0.071 (\pm 0.007)	0.18 (\pm 0.03)	0.59 (\pm 0.08)

Global rhizarian NPP was derived from mean estimates (\pm s.e.m., computed from variation in rhizarian abundance in our dataset) assuming a total ocean surface of 3.61×10^{14} m² of which nearly 56% is considered oligotrophic (that is, 2.04×10^{14} m²; ref. 6). The rhizarian NPP contribution to global and oligotrophic regions was calculated from total NPP estimates of 48.5 Pg C and 11 Pg C per year, respectively⁴⁴. Contribution in the oligotrophic regions was estimated by extracting mean rhizarian NPP estimates in sampling stations where the Chl_a_{sat} was <0.1 mg m⁻³ (ref. 44). Contribution to the >2- μ m size fraction assumed that pico-phytoplankton contributed up to 70% in oligotrophic regions⁴⁵.

Musashi-2 attenuates AHR signalling to expand human haematopoietic stem cells

Stefan Rentas¹, Nicholas T. Holzapfel^{1*}, Muluken S. Belew^{1*}, Gabriel A. Pratt^{2,3*}, Veronique Voisin⁴, Brian T. Wilhelm⁵, Gary D. Bader⁴, Gene W. Yeo^{2,3,6} & Kristin J. Hope¹

Umbilical cord blood-derived haematopoietic stem cells (HSCs) are essential for many life-saving regenerative therapies. However, despite their advantages for transplantation, their clinical use is restricted because HSCs in cord blood are found only in small numbers¹. Small molecules that enhance haematopoietic stem and progenitor cell (HSPC) expansion in culture have been identified^{2,3}, but in many cases their mechanisms of action or the nature of the pathways they impinge on are poorly understood. A greater understanding of the molecular circuitry that underpins the self-renewal of human HSCs will facilitate the development of targeted strategies that expand HSCs for regenerative therapies. Whereas transcription factor networks have been shown to influence the self-renewal and lineage decisions of human HSCs^{4,5}, the post-transcriptional mechanisms that guide HSC fate have not been closely investigated. Here we show that overexpression of the RNA-binding protein Musashi-2 (MSI2) induces multiple pro-self-renewal phenotypes, including a 17-fold increase in short-term repopulating cells and a net 23-fold *ex vivo* expansion of long-term repopulating HSCs. By performing a global analysis of MSI2–RNA interactions, we show that MSI2 directly attenuates aryl hydrocarbon receptor (AHR) signalling through post-transcriptional downregulation of canonical AHR pathway components in cord blood HSPCs. Our study gives mechanistic insight into RNA networks controlled by RNA-binding proteins that underlie self-renewal and provides evidence that manipulating such networks *ex vivo* can enhance the regenerative potential of human HSCs.

Control of translation by RNA-binding proteins in human HSCs and its potential to regulate HSC self-renewal remain underexplored. MSI2 is known to regulate mouse HSCs^{6–8} and has been predicted to influence mRNA translation⁹, so we investigated the role of MSI2 in post-transcriptionally controlling self-renewal of human HSPCs. The expression of *MSI2* mRNA was elevated in primitive cord blood HSPCs and decreased during differentiation, whereas the *MSI2* paralogue, *MSI1*, was not expressed (Extended Data Fig. 1a–f). Lentiviral overexpression of MSI2 resulted in a 1.5-fold increase in colony-forming units (CFU) relative to control cells, principally due to a 3.7-fold increase in the most primitive CFU-granulocyte erythrocyte monocyte megakaryocyte (GEMM) colony type (Extended Data Fig. 2a and Fig. 1a). Remarkably, 100% of MSI2-overexpressing CFU-GEMMs generated secondary colonies compared to only 40% of control CFU-GEMMs. In addition, MSI2 overexpression yielded three times as many colonies per re-seeded CFU-GEMM (Fig. 1b, c and Extended Data Fig. 2b). During *in vitro* culture, MSI2-overexpressing cells were 2.3- and 6-fold more abundant than control cells at the 7- and 21-day time points, respectively (Extended Data Fig. 2c, d). Moreover, after 7 days in culture, MSI2-overexpressing cells showed a cumulative 9.3-fold increase in colony-forming cells in the absence of changes in cell cycling or

death (Extended Data Fig. 2e–h). Together, our data demonstrate that enforced expression of MSI2 has potent self-renewal-inducing effects on early progenitors and promotes their *in vitro* expansion.

Short-term repopulating cells (STRCs) produce a transient multilineage graft in non-obese diabetic (NOD)/SCID *Il2r^{null}* (NSG) mice¹⁰, and in patients these cells reconstitute granulocytes and platelets that are essential for preventing post-transplantation infection and bleeding¹. MSI2-overexpressing STRCs yielded 1.8-fold more primitive CD34⁺ cells post-infection and a 17-fold increase in functional STRCs relative to control STRCs, as determined by limiting dilution analysis (LDA) of human chimaerism in mice 3 weeks after transplantation (Fig. 1d–f and Extended Data Fig. 3a, b). Furthermore, at a protracted engraftment

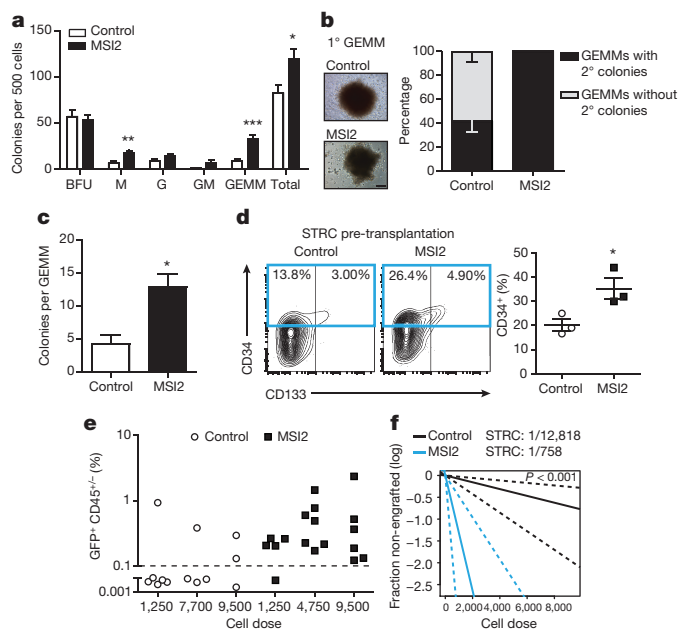


Figure 1 | MSI2 overexpression enhances *in vitro* cord blood progenitor activity and increases the number of STRCs. **a**, CFU output from transduced Lin[−] cord blood ($n = 9$ control and 10 MSI2-overexpressing (MSI2) cultures from 5 experiments). **b**, CFU-GEMM secondary CFU replating potential ($n = 24$ control and 30 MSI2-overexpressing cultures from 2 experiments) and images of primary GEMMs (scale bar, 200 μ m). **c**, Number of secondary colonies per replated CFU-GEMM from **b**. **d**, CD34 expression in STRCs before transplantation ($n = 3$ experiments). **e**, Human chimaerism at 3 weeks in mice transplanted with varying doses of transduced STRCs. Dashed line indicates engraftment cut-off ($n = 3$ experiments). **f**, STRC frequency determined by LDA from **e**. Dashed lines indicate 95% confidence intervals. Data shown as mean \pm s.e.m. Unpaired *t*-test, * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

¹Department of Biochemistry and Biomedical Sciences, Stem Cell and Cancer Research Institute, McMaster University, Hamilton, Ontario L8S 4K1, Canada. ²Department of Cellular and Molecular Medicine, Institute for Genomic Medicine, University of California, San Diego, La Jolla, California 92037, USA. ³Bioinformatics Graduate Program, University of California, San Diego, La Jolla, California 92037, USA. ⁴The Donnelly Centre, University of Toronto, Toronto, Ontario M5S 3E1, Canada. ⁵Institute for Research in Immunology and Cancer, University of Montreal, Montreal, Quebec H3C 3J7, Canada. ⁶Department of Physiology, National University of Singapore and Molecular Engineering Laboratory, A*STAR, Singapore 138632, Singapore.

*These authors contributed equally to this work.

readout time of 6.5 weeks at non-limiting transplant doses, 100% of mice transplanted with MSI2-overexpressing STRCs were engrafted compared to only 50% of mice transplanted with control STRCs, indicating that MSI2 overexpression extended the duration of STRC-mediated engraftment (Extended Data Fig. 3c).

We next explored the effect of short hairpin (sh)RNA-induced MSI2 knockdown on HSPC function. MSI2 knockdown did not alter the clonogenic potential of HSPCs but did decrease CFU replating threefold (Extended Data Fig. 4a–c). In more primitive culture-initiating cells, MSI2 knockdown significantly decreased cell numbers over culture (Extended Data Fig. 4d, e) independent of increased death or altered cell cycling (data not shown). Upon transplantation, engrafted MSI2 knockdown GFP⁺ cells showed no evidence of lineage skewing, but the frequency of cells was markedly reduced relative to the percentage of GFP⁺ cells initially transplanted (Extended Data Fig. 4f–h). Combined, our *in vitro* and *in vivo* data show that MSI2 knockdown reduces self-renewal in early progenitors and HSCs.

To characterize the earliest transcriptional changes induced by modulating MSI2 expression, we performed RNA sequencing (RNA-seq) on CD34⁺ MSI2-overexpressing and knockdown cells immediately after transduction (Supplementary Tables 1 and 2). MSI2 overexpression-induced transcriptional changes showed an inverse correlation with those induced by MSI2 knockdown, suggesting that overexpression and knockdown had opposite effects (Extended Data Fig. 5a). When compared to transcriptome data from 38 human haematopoietic cell subpopulations⁴, genes that were significantly upregulated by MSI2 overexpression and downregulated upon MSI2 knockdown were exclusively enriched in those highly expressed in HSCs and other primitive CD34⁺ populations (Extended Data Fig. 5b).

As MSI2 overexpression conferred an HSC gene expression program, we hypothesized that it could facilitate HSC expansion *ex vivo*. MSI2 overexpression induced a fourfold increase in CD34⁺CD133⁺ phenotypic HSCs relative to control cells after 7 days of culture (Fig. 2a). We next performed an LDA to define functional HSC frequency before (day 3 post-transduction, D3) and after 7 days of *ex vivo* culture (day 10, D10; Extended Data Fig. 6a). Mice transplanted with HSCs on D3 displayed no altered engraftment as a result of MSI2 overexpression; however, recipients of MSI2-overexpressing D10-expanded cells displayed multiple phenotypes of enhanced reconstitution relative to recipients of control cells, including a twofold increase in bone marrow GFP⁺ levels without changes to lineage output, an increase in the proportion of GFP⁺ cells within the human graft relative to pre-transplant D10 levels, an increase in GFP mean fluorescence intensity and enrichment of CD34 expression in GFP^{high} cells (Fig. 2b, c and Extended Data Fig. 6b–h). As the lentiviral construct design ensures that levels of GFP mirror those of MSI2, these findings indicate that high levels of MSI2 impart enhanced competitiveness and are conducive to *in vivo* HSPC activity. Importantly, D10 MSI2-overexpressing cultures contained more CD34⁺CD133⁺ cells before transplantation than did control cultures (Extended Data Fig. 6i) and, accordingly, the HSC frequency in D10 MSI2-overexpressing cultures was increased twofold relative to that in D3 MSI2-overexpressing cultures. By contrast, control cultures displayed a threefold decrease in HSC frequency. These results demonstrate that MSI2 overexpression *ex vivo* facilitated a net sixfold increase in HSC frequency relative to control cultures (Fig. 2g, h and Supplementary Tables 3, 4).

Secondary LDA transplants were performed to explore fully the effects of MSI2 overexpression and culturing on self-renewal and long-term HSCs (LT-HSCs). Robust engraftment with MSI2-overexpressing cells did not induce altered myelo-lymphopoiesis or leukaemic development (Fig. 2e). Secondary LDA measurements revealed that the percentage of GFP⁺ cells in the bone marrow was 4.6-fold higher after transplantation of MSI2-overexpressing cells than after control transplantation, and LT-HSC frequency was 3.5-fold higher (Fig. 2d, f and Supplementary Table 5). The increase in LT-HSC frequency corresponds to MSI2-overexpressing GFP⁺ HSCs having expanded in

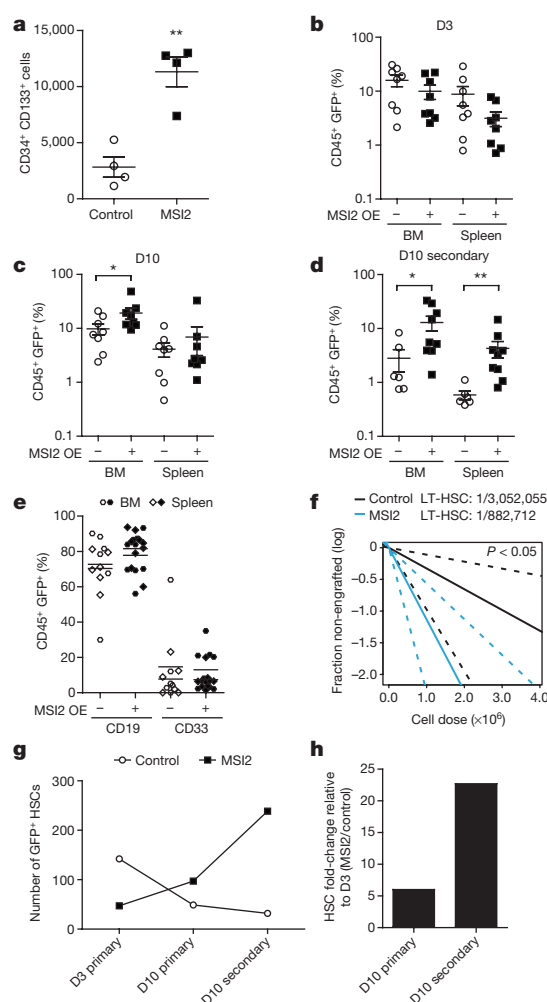


Figure 2 | MSI2 overexpression expands LT-HSCs in *ex vivo* culture.

a, Transduced CD34⁺CD133⁺ cells after 1 week of culture ($n = 4$ experiments, unpaired *t*-test). **b–d**, CD45⁺GFP⁺ engraftment in mice that received the highest two cell doses at D3 and D10 ($n = 8$ mice for both control and MSI2-overexpressing (MSI2 OE) cells) and the highest three cell doses of D10 secondary cells ($n = 6$ mice for control and 9 for MSI2-overexpressing cells, Mann–Whitney test). BM, bone marrow. **e**, Myelo-lymphopoiesis in mice that received D10 secondary cells. **f**, Multi-lineage LT-HSC frequency in bone marrow cells from mice that received D10 primary cells. Dashed lines indicate 95% confidence intervals. **g**, Numbers of GFP⁺ HSCs as evaluated by LDA. **h**, Cumulative fold change in MSI2-overexpressing HSCs. Data shown as mean \pm s.e.m. * $P < 0.05$; ** $P < 0.01$.

primary mice 2.4-fold over input as compared to a 1.5-fold decrease for control HSCs (Fig. 2g). The level of *in vivo* expansion induced by MSI2 overexpression reflects the behaviour of uncultured HSCs, which undergo similarly controlled expansion during passage in mice^{3,11,12}. Finally, when accounting for the total change in GFP⁺ HSCs upon *ex vivo* culture, MSI2 overexpression induced a cumulative 23-fold expansion of secondary LT-HSCs relative to control (Fig. 2g, h), indicating that elevated MSI2 expression provides a considerable self-renewal advantage to functional HSCs during *ex vivo* culture.

To gain mechanistic insight into this process, we examined genes that were differentially expressed in MSI2-overexpressing cells and found that the gene encoding cytochrome P450 1B1 oxidase (CYP1B1), an effector of AHR signalling¹³, was among the most repressed (Supplementary Table 1). Pathway analysis revealed that many predicted targets of AHR were enriched in the gene sets that were downregulated by MSI2 overexpression (Fig. 3a) and upregulated by MSI2 knockdown (Extended Data Fig. 7a, b). Binding of

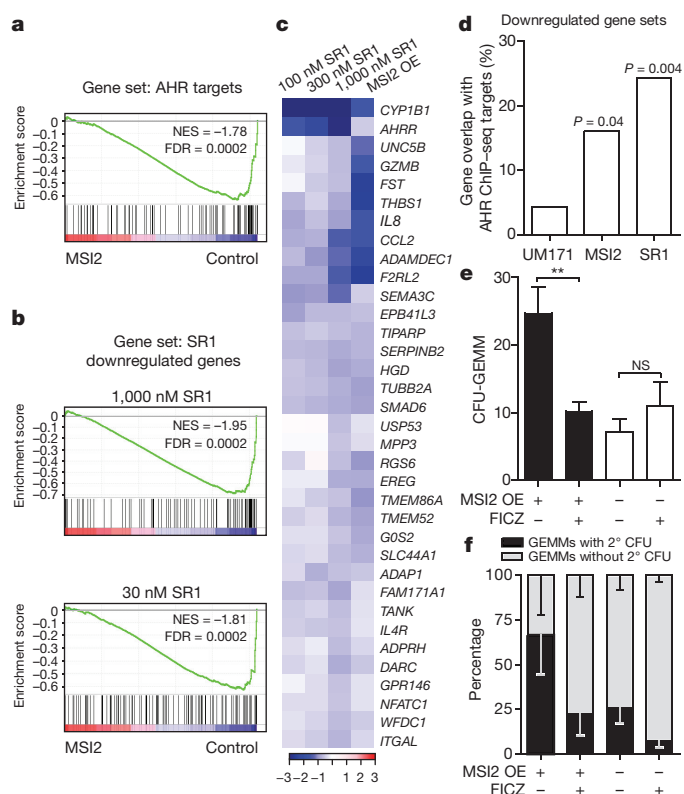


Figure 3 | MSI2 overexpression in human HSPCs attenuates AHR signalling. **a**, Predicted AHR targets compared by GSEA to genes downregulated by MSI2 overexpression. **b**, GSEA of gene sets downregulated by SR1 and MSI2 overexpression. **c**, log fold-change in expression of GSEA leading edge genes downregulated by MSI2 overexpression or SR1. **d**, Percentage of gene overlap between gene sets downregulated by UM171, SR1 or MSI2 overexpression and AHR targets identified by ChIP-seq. **e**, Number of CFU-GEMMs generated from transduced cells grown in CFU medium containing FICZ or dimethylsulfoxide (DMSO) ($n = 3$ experiments). **f**, CFU-GEMMs from **e** replated into CFU assays containing FICZ or DMSO ($n = 30$ control and 29 MSI2 overexpressing CFU-GEMMs per treatment). Data are presented as mean \pm s.e.m. Unpaired t -test, $**P < 0.01$. NS, not significant.

the nuclear receptor transcription factor AHR to the small molecule StemRegenin 1 (SR1) inhibits AHR target gene activation and leads to the expansion of human HSPCs in culture². Gene set enrichment analysis (GSEA) showed that genes that were downregulated by MSI2 overexpression significantly matched those that were downregulated by SR1 in an SR1 dose-dependent manner (Fig. 3b, c), whereas MSI2 knockdown induced the opposite expression profile (Extended Data Fig. 7c, d). We next examined the overlap between genes that were downregulated by MSI2 overexpression and AHR targets identified by chromatin-immunoprecipitation followed by sequencing (ChIP-seq)¹⁴. This comparison was extended to genes downregulated upon treatment with UM171, which expands HSPCs independently of AHR³. Direct transcriptional targets of AHR were enriched by 3.8- and 5.6-fold in the gene sets that were downregulated by MSI2 overexpression and SR1, respectively, compared to UM171; this overrepresentation was maintained for predicted AHR targets and suggests that MSI2 overexpression expands HSPCs by attenuating AHR signalling (Fig. 3d and Extended Data Fig. 7e). Furthermore, SR1 treatment increased the percentage of CD34⁺ cells eightfold in control cultures but only fourfold in MSI2-overexpressing cultures (Extended Data Fig. 8a, b), a finding that suggests that SR1 and MSI2 overexpression act redundantly on HSPCs via the same pathway.

To elucidate further the connection between MSI2 and AHR, MSI2-overexpressing and control cultures were treated with the AHR agonist 6-formylindolo(3,2-b)carbazole (FICZ).

Treatment of MSI2-overexpressing cells with FICZ induced canonical AHR targets, showing that these cells remain competent for AHR activation (Extended Data Fig. 8c). FICZ induced a marked reversal of the MSI2 overexpression-mediated increases in primary CFU-GEMMs and their replating capacities (Fig. 3e, f). Furthermore, FICZ-treated MSI2-overexpressing cultures displayed greater losses of phenotypic HSPCs compared to treated controls, which showed no change (Extended Data Fig. 8d, e). Together, these results show that agonist-induced restoration of AHR activity reduces the self-renewal-promoting effects of MSI2 overexpression and strongly supports the idea that MSI2 overexpression promotes HSPC expansion through downregulation of AHR signalling.

To identify key RNA targets that underlie MSI2 function, we analysed global MSI2 protein–RNA interactions using cross-linking immunoprecipitation followed by sequencing (CLIP-seq)¹⁵ (Extended Data Fig. 9a, b). Replicates were highly correlated via gene RPKMs (reads per kilobase of transcript per million mapped reads) and 5,552 protein-coding genes were bound in both replicates (Extended Data Fig. 9c and Fig. 4a, b). Within the top 40% of reproducible clusters, MSI2 bound predominantly to the 3' untranslated regions (3'UTRs) of mature mRNAs (Fig. 4c). Importantly, 9% of annotated protein-coding gene mRNAs were reproducible MSI2 targets, compared to 0.2% of long non-coding RNAs (Extended Data Fig. 9d), suggesting that MSI2 controls the stability or translation of coding mRNAs. Motif analysis identified a consensus pentamer (U/G)UAGU resembling the known mouse Msi1-binding sequence^{9,16} within binding sites in all genic regions; additionally, MSI2-binding sites were generally significantly more conserved than background and tended to occur after the stop codon (Fig. 4d and Extended Data Fig. 9e–h). The presence of MSI2 binding sites within Msi1 targets¹⁶ across species indicates that Musashi proteins may bind the same genes through 3'UTR-embedded motifs (Extended Data Fig. 9i). Finally, target gene ontology analysis revealed 186 biological processes categories (Supplementary Table 6), among the most significant of which were electron transport, oestrogen receptor signalling regulation and metabolism of small molecules, all processes known to be transcriptionally influenced by AHR signalling¹⁷.

Among the top 2% of enriched CLIP-seq targets (Supplementary Table 7) were the 3'UTRs of the genes for two AHR pathway components: heat shock protein 90 (HSP90) and CYP1B1. Each exhibited multiple MSI2-binding motifs correlating with overlapping clusters of CLIP-seq reads (Fig. 4e and Extended Data Fig. 10a). To investigate the ability of MSI2 to post-transcriptionally regulate these genes during HSPC expansion, we looked for instances of uncoupled transcript and protein expression. HSP90 displayed uncoupling of transcript (1.6-fold up) and protein (1.6-fold down) expression early in culture, but after 7 days showed further upregulated transcript expression (2.5-fold) and variable protein levels (Fig. 4f and Extended Data Fig. 10b). As AHR–HSP90 binding is essential for ligand-dependent transcriptional activity¹³, downregulation of HSP90 protein at the outset of HSPC culture would be expected to reduce latent AHR complex formation and attenuate AHR signalling (Fig. 3a). Indeed, CYP1B1 transcript and protein expression displayed twofold reductions early in culture, consistent with decreased AHR pathway activity; however, at day 7, CYP1B1 transcripts were upregulated 1.7-fold and uncoupled from protein expression, which was downregulated twofold (Fig. 4g and Extended Data Fig. 10c). To test whether MSI2 directly mediates post-transcriptional repression of these targets, the 3'UTRs of CYP1B1 and HSP90 were coupled to luciferase. MSI2 overexpression induced significant reductions in luciferase signal from both reporters, and this effect was mitigated when the core CLIP-seq-identified UAG motifs were mutated (Extended Data Fig. 10d, e). As MSI2 overexpression-mediated post-transcriptional downregulation of the AHR pathway converged on CYP1B1 protein repression throughout culture, we explored the effects on HSPCs of inhibiting CYP1B1 independently with (E)-2,3',4,5'-tetramethoxystilbene (TMS). During culture, TMS

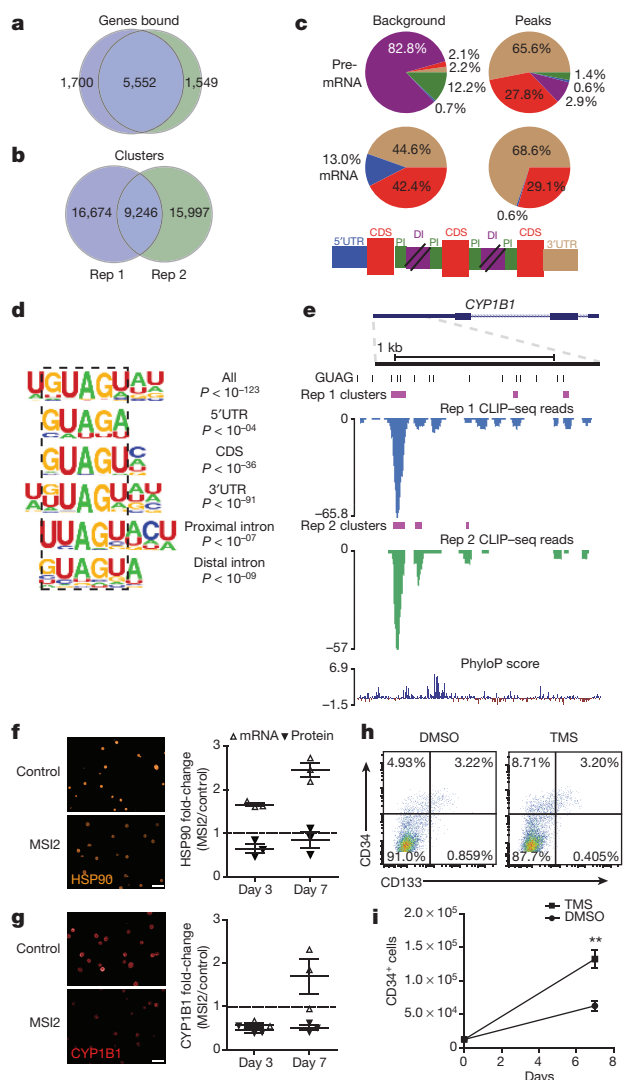


Figure 4 | MSI2 overexpression post-transcriptionally downregulates AHR pathway components. **a**, Overlap between MSI2 target genes from separate CLIP-seq experiments. **b**, Statistically significant overlap ($P < 0.0001$, hypergeometric test) of clusters between the replicates. **c**, Percentage of CLIP-seq clusters in different genic regions. **d**, Consensus motifs within MSI2 clusters in different genic regions. P values presented for the top 40% of clusters. **e**, CLIP-seq reads (blue, replicate 1; green, replicate 2) and clusters (purple) mapped to the 3'UTR of *CYP1B1*. Matches to the GUAG motif are shown in black. **f**, **g**, Immunofluorescence for HSP90 and CYP1B1 3 days after transduction and summary of fold-changes in HSP90 and CYP1B1 protein and transcript levels with MSI2 overexpression at 3 and 7 days after transduction (scale bar, 20 μ m; dotted line indicates no change; $n = 3$ experiments). **h**, HSPC marker expression by CD34⁺ cells treated with TMS for 10 days. **i**, Absolute CD34⁺ cell number with TMS ($n = 4$ experiments). Data are presented as mean \pm s.e.m. Unpaired t -test, $**P < 0.01$.

increased the frequency and total numbers of CD34⁺ cells by 1.5-fold and 2-fold, respectively (Fig. 4h, i), phenocopying the effects of MSI2 overexpression. Finally, overexpression of both CYP1B1 lacking its 3'UTR and MSI2 decreased secondary CFU-GEMM replating efficiency (Extended Data Fig. 10f, g); this suggests that CYP1B1, while typically used to report AHR signalling, itself promotes HSPC differentiation.

Our work identifies MSI2 as an important mediator of human HSPC self-renewal and *ex vivo* expansion that acts by coordinating the post-transcriptional regulation of proteins belonging to a shared self-renewal regulatory pathway (Extended Data Fig. 10h). We envision that manipulation of the post-transcriptional circuitry controlled by RNA-binding proteins will provide a novel and powerful means by

which to enhance the regenerative potential of not only human HSCs but also other stem-cell types.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 June 2015; accepted 15 March 2016.

- Miller, P. H., Knapp, D. J. & Eaves, C. J. Heterogeneity in hematopoietic stem cell populations: implications for transplantation. *Curr. Opin. Hematol.* **20**, 257–264 (2013).
- Boitano, A. E. *et al.* Aryl hydrocarbon receptor antagonists promote the expansion of human hematopoietic stem cells. *Science* **329**, 1345–1348 (2010).
- Fares, I. *et al.* Pyrimidoindole derivatives are agonists of human hematopoietic stem cell self-renewal. *Science* **345**, 1509–1512 (2014).
- Novershtern, N. *et al.* Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* **144**, 296–309 (2011).
- Laurenti, E. *et al.* The transcriptional architecture of early human hematopoiesis identifies multilevel control of lymphoid commitment. *Nature Immunol.* **14**, 756–763 (2013).
- Hope, K. J. *et al.* An RNAi screen identifies Msi2 and Prox1 as having opposite roles in the regulation of hematopoietic stem cell activity. *Cell Stem Cell* **7**, 101–113 (2010).
- de Andrés-Aguayo, L. *et al.* Musashi 2 is a regulator of the HSC compartment identified by a retroviral insertion screen and knockout mice. *Blood* **118**, 554–564 (2011).
- Park, S. M. *et al.* Musashi-2 controls cell fate, lineage bias, and TGF- β signaling in HSCs. *J. Exp. Med.* **211**, 71–87 (2014).
- Ohyama, T. *et al.* Structure of Musashi1 in a complex with target RNA: the role of aromatic stacking interactions. *Nucleic Acids Res.* **40**, 3218–3231 (2012).
- Glimm, H. *et al.* Previously undetected human hematopoietic cell populations with short-term repopulating activity selectively engraft NOD/SCID- β 2 microglobulin-null mice. *J. Clin. Invest.* **107**, 199–206 (2001).
- Cashman, J. D. & Eaves, C. J. Human growth factor-enhanced regeneration of transplantable human hematopoietic stem cells in nonobese diabetic/severe combined immunodeficient mice. *Blood* **93**, 481–487 (1999).
- Holyoake, T. L., Nicolini, F. E. & Eaves, C. J. Functional differences between transplantable human hematopoietic stem cells from fetal liver, cord blood, and adult marrow. *Exp. Hematol.* **27**, 1418–1427 (1999).
- Mimura, J. & Fujii-Kuriyama, Y. Functional role of AhR in the expression of toxic effects by TCDD. *Biochim. Biophys. Acta* **1619**, 263–268 (2003).
- Lo, R. & Matthews, J. High-resolution genome-wide mapping of AHR and ARNT binding sites by ChIP-seq. *Toxicol. Sci.* **130**, 349–361 (2012).
- Yeo, G. W. *et al.* An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nature Struct. Mol. Biol.* **16**, 130–137 (2009).
- Katz, Y. *et al.* Musashi proteins are post-transcriptional regulators of the epithelial-luminal cell state. *Elife* **3**, e03915 (2014).
- Tijet, N. *et al.* Aryl hydrocarbon receptor regulates distinct dioxin-dependent and dioxin-independent gene batteries. *Mol. Pharmacol.* **69**, 140–153 (2006).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank E. Lechman and P. Van Galen for experimental advice and for providing H1 and pSMALB vectors. The MA overexpression vector was a gift from L. Naldini. We also thank the SCC-RI core flow cytometry staff, the Obstetrics and Gynecology Unit at McMaster Children's Hospital for cord blood, B. Doble and M. Bhatia for critical assessment of this work and all members of the Hope laboratory for experimental support and advice. This work was supported by an Ontario Institute for Cancer Research New Investigator Award (IA-033), an Ontario Institute for Cancer Research Cancer Stem Cell Program Team Grant (PCSC.005) and a Canadian Institutes of Health Research (MOP-126030) grant to K.J.H. N.T.H. was supported in part by a CIHR MD/PhD Studentship. M.S.B. was supported by an NSERC Alexander Graham Bell Doctoral Fellowship. S.R. is supported by a Canadian Blood Services Graduate Fellowship and Health Canada. The views expressed herein do not necessarily represent the view of the federal government of Canada. This work was partially supported by grants from the National Institute of Health (HG004659 and NS075449) and the California Institute of Regenerative Medicine (RB3-05219) to G.W.Y. G.P. was supported by a National Science Graduate Fellowship. G.W.Y. is an Alfred P. Sloan Research Fellow. We thank the UCSD Institute for Genomic Medicine's Genomics Center for providing access to high-throughput sequencing facilities.

Author Contributions S.R. designed and performed experiments, analysed data and wrote the manuscript. N.T.H. constructed CLIP-seq libraries. M.S.B. helped perform cord blood experiments. G.A.P. and G.W.Y. advised on CLIP-seq library construction, performed CLIP-seq bioinformatic analyses and wrote the manuscript. B.T.W. performed RNA-seq analyses. V.V. and G.D.B. performed RNA-seq bioinformatic analyses. K.J.H. conceived the project, supervised the study, analysed data, interpreted results and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to G.W.Y. (geneyeo@ucsd.edu) or K.J.H. (kristin@mcmaster.ca).

METHODS

Mice. NOD/SCID *Il2rg*^{null} mice (Jackson Laboratory) were bred and maintained in the Stem Cell Unit animal barrier facility at McMaster University. All procedures were approved by the Animal Research Ethics Board at McMaster University.

Isolation of primitive human haematopoietic cells and flow cytometry. All patient samples were obtained with informed consent and with the approval of local human subject research ethics boards at McMaster University. Human umbilical cord blood mononuclear cells were collected by centrifugation with Ficoll-Paque Plus (GE), followed by red blood cell lysis with ammonium chloride (StemCell Technologies). Cells were then incubated with a cocktail of lineage-specific antibodies (CD2, CD3, CD11b, CD11c, CD14, CD16, CD19, CD24, CD56, CD61, CD66b, and GlyA; StemCell Technologies) for negative selection of Lin[−] cells using an EasySep immunomagnetic column (StemCell Technologies). Live cells were discriminated on the basis of cell size, granularity and, as needed, absence of viability dye 7-AAD (BD Biosciences) uptake. All flow cytometry analysis was performed using a BD LSR II instrument (BD Biosciences). Data acquisition was conducted using BD FACSDiva software (BD Biosciences) and analysis was performed using FlowJo software (Tree Star).

HSPC sorting and qRT-PCR analysis. To quantify *MSI2* expression in human HSPCs, Lin[−] cord blood cells were stained with the appropriate antibody combinations to resolve HSC (CD34⁺ CD38[−] CD45RA[−] CD90⁺), MPP (CD34⁺ CD38[−] CD45RA[−] CD90[−]), CMP (CD34⁺ CD38⁺ CD71[−]) and EP (CD34⁺ CD38⁺ CD71⁺) fractions as similarly described previously^{18,19} with all antibodies from BD Biosciences: CD45RA (HI100), CD90 (5E10), CD34 (581), CD38 (HB7) and CD71 (M-A712). Cell viability was assessed using the viability dye 7AAD (BD Biosciences). All cell subsets were isolated using a BD FACSAria II cell sorter (BD Biosciences) or a MoFlo XDP cell sorter (Beckman Coulter). HemaExplorer²⁰ analysis was used to confirm *MSI2* expression in human HSPCs and across the hierarchy. For all qRT-PCR determinations total cellular RNA was isolated with TRIzol LS reagent according to the manufacturer's instructions (Invitrogen) and cDNA was synthesized using the qScript cDNA Synthesis Kit (Quanta Biosciences). qRT-PCR was done in triplicate with PerfeCTa qPCR SuperMix Low ROX (Quanta Biosciences) with gene-specific probes (Universal Probe Library (UPL), Roche) and primers: *MSI2* UPL-26, F-GGCAGCAAGAGGATCAGG, R-CCGTAGAGATCGGCGACA; *HSP90* UPL-46, F-GGGCAACACCTCTACAAGGA, R-CTTGGGTCTGGGTTTCCTC; *CYP11B1* UPL-20, F-ACGTACCGGCCACTATCACT, R-CTCGAGTCTGCACATCAGGA; *GAPDH* UPL-60, F-AGCCACATCGCTCAGACAC, R-GCCCAA TACGACCAAATCC; *ACTB* (UPL Set Reference Gene Assays, Roche). The mRNA content of samples compared by qRT-PCR was normalized based on the amplification of *GAPDH* or *ACTB*.

Lentivirus production and western blot validation. *MSI2* shRNAs were designed with the Dharmacon algorithm (<http://www.dharmacon.com>). Predicted sequences were synthesized as complementary oligonucleotides, annealed and cloned downstream of the H1 promoter of the modified cppt-PGK-EGFP-IRES-PAC-WPRE lentiviral expression vector¹⁸. Sequences for the *MSI2* targeting and control RFP targeting shRNAs were as follows: sh*MSI2*, 5'-GAGAGATCCCACACTACGAAA-3'; shRFP, 5'-GTGGGAGCGCGTGATGAAC-3'. Human *MSI2* cDNA (BC001526; Open Biosystems) was subcloned into the MA bi-directional lentiviral expression vector²¹. Human *CYP11B1* cDNA (BC012049; Open Biosystems) was cloned in to psMALB²². All lentiviruses were prepared by transient transfection of 293FT (Invitrogen) cells with pMD2.G and psPAX2 packaging plasmids (Addgene) to create VSV-G pseudotyped lentiviral particles. All viral preparations were titrated on HeLa cells before use on cord blood. Standard SDS-PAGE and western blotting procedures were performed to validate the effects of knockdown on transduced NB4 cells (DSMZ) and overexpression on 293FT cells. Immunoblotting was performed with anti-*MSI2* rabbit monoclonal IgG (EP1305Y, Epitomics) and β -actin mouse monoclonal IgG (ACTB11B7, Santa Cruz Biotechnology) antibodies. Secondary antibodies used were IRDye 680 goat anti-rabbit IgG and IRDye 800 goat anti-mouse IgG (LI-COR). 293FT and NB4 cell lines tested negative for mycoplasma. NB4 cells were authenticated by ATRA treatment before use.

Cord blood transduction. Cord blood transductions were conducted as described previously^{18,23}. Briefly, thawed Lin[−] cord blood or flow-sorted Lin[−] CD34⁺ CD38[−] or Lin[−] CD34⁺ CD38⁺ cells were prestimulated for 8–12 h in StemSpan medium (StemCell Technologies) supplemented with growth factors interleukin 6 (IL-6; 20 ng ml^{−1}, Peprotech), stem cell factor (SCF; 100 ng ml^{−1}, R&D Systems), Flt3 ligand (FLT3-L; 100 ng ml^{−1}, R&D Systems) and thrombopoietin (TPO; 20 ng ml^{−1}, Peprotech). Lentivirus was then added in the same medium at a multiplicity of infection of 30–100 for 24 h. Cells were then given 2 days after transduction before use in *in vitro* or *in vivo* assays. For *in vitro* cord blood studies biological (experimental) replicates were performed with three independent cord blood samples.

Clonogenic progenitor assays. Human clonogenic progenitor cell assays were done in semi-solid methylcellulose medium (Methocult H4434; StemCell

Technologies) with flow-sorted GFP⁺ cells post transduction (500 cells per ml) or from day seven cultured transduced cells (12,000 cells per ml). Colony counts were carried out after 14 days of incubation. CFU-GEMMs can seed secondary colonies owing to their limited self-renewal potential²⁴. Replating of *MSI2*-overexpressing and control CFU-GEMMs for secondary CFU analysis was performed by picking single CFU-GEMMs at day 14 and disassociating colonies by vortexing. Cells were spun and resuspended in fresh methocult, mixed with a blunt-ended needle and syringe, and then plated into single wells of a 24-well plate. Secondary CFU analysis for sh*MSI2*- and shControl-expressing cells was performed by harvesting total colony growth from a single dish (as nearly equivalent numbers of CFU-GEMMs were present in each dish), resuspending cells in fresh methocult by mixing vigorously with a blunt-ended needle and syringe and then plating into replicate 35-mm tissue culture dishes. In both protocols, secondary colony counts were done following incubation for 10 days. For primary and secondary colony forming assays performed with the AHR agonist FICZ (Santa Cruz Biotechnology), 200 nM FICZ or 0.1% DMSO was added directly to H4434 methocult medium. Two-way ANOVA analysis was performed to compare secondary CFU output and FICZ treatment for *MSI2*-overexpressing or control conditions. Colonies were imaged with a Q-Colour3 digital camera (Olympus) mounted to an Olympus IX5 microscope with a 10 \times objective lens. Image-Pro Plus imaging software (Media Cybernetics) was used to acquire pictures and subsequent image processing was performed with ImageJ software (NIH).

Lin[−] cord blood and Lin[−] CD34⁺ suspension cultures. Transduced human Lin[−] cord blood cells were sorted for GFP expression and seeded at a density of 10⁵ cells per ml in IMDM 10% FBS supplemented with human growth factors IL-6 (10 ng ml^{−1}), SCF (50 ng ml^{−1}), FLT3-L (50 ng ml^{−1}), and TPO (20 ng ml^{−1}) as previously described²⁵. To generate growth curves, every seven days cells were counted, washed, and resuspended in fresh medium with growth factors at a density of 10⁵ cells per ml. Cells from suspension cultures were also used in clonogenic progenitor, cell cycle and apoptosis assays. Experiments performed on transduced Lin[−] CD34⁺ cord blood cells used serum-free conditions as described in the cord blood transduction subsection of Methods. For *in vitro* cord blood studies, biological (experimental) replicates were performed with three independent cord blood samples.

Cell cycle and apoptosis assays. Cell cycle progression was monitored with the addition of BrdU to day 10 suspension cultures at a final concentration of 10 μ M. After 3 h of incubation, cells were assayed with the BrdU Flow Kit (BD Biosciences) according to the manufacturer's protocol. Cell proliferation and quiescence were measured using Ki67 (BD Bioscience) and Hoechst 33342 (Sigma) on day 4 suspension cultures after fixing and permeabilizing cells with the Cytofix/Cytoperm kit (BD Biosciences). For apoptosis analysis, Annexin V (Invitrogen) and 7-AAD (BD Bioscience) staining of day 7 suspension cultures was performed according to the manufacturer's protocol.

Intracellular flow cytometry. Lin[−] cord blood cells were initially stained with anti-CD34 PE (581) and anti-CD38 APC (HB7) antibodies (BD Biosciences) then fixed with the Cytofix/Cytoperm kit (BD Biosciences) according to the manufacturer's instructions. Fixed and permeabilized cells were immunostained with anti-*MSI2* rabbit monoclonal IgG antibody (EP1305Y, Abcam) and detected by Alexa-488 goat anti-rabbit IgG antibody (Invitrogen).

RNA-seq data processing. CD34⁺ cells were transduced with an *MSI2*-overexpression or *MSI2*-knockdown lentivirus along with their corresponding controls and sorted for GFP expression 3 days later. Transductions for *MSI2* overexpression or knockdown were each performed on two independent cord blood samples. Total RNA from transduced cells (>1 \times 10⁵) was isolated using TRIzol LS as recommended by the manufacturer (Invitrogen), and then further purified using RNeasy columns (Qiagen). Sample quality was assessed using Bioanalyzer RNA Nano chips (Agilent). Paired-end, barcoded RNA-seq sequencing libraries were then generated using the TruSeq RNA Sample Prep Kit (v2) (Illumina) following the manufacturer's protocols starting from 1 μ g total RNA. The quality of library generation was then assessed using a Bioanalyzer platform (Agilent) and Illumina MiSeq-QC run was performed or quantified by qPCR using KAPA quantification kit (KAPA Biosystems). Sequencing was performed using an Illumina HiSeq2000 using TruSeq SBS v3 chemistry at the Institute for Research in Immunology and Cancer's Genomics Platform (University of Montreal) with cluster density targeted at 750,000 clusters per mm² and paired-end 2 \times 100-bp read lengths. For each sample, 90–95 million reads were produced and mapped to the hg19 (GRCh37) human genome assembly using CASAVA (version 1.8). Read counts generated by CASAVA were processed in EdgeR (edgeR_3.12.0, R 3.2.2) using TMM normalization, paired design, and estimation of differential expression using a generalized linear model (glmFit). The false discovery rate (FDR) was calculated from the output *P* values using the Benjamini–Hochberg method. The fold change of logarithm of base 2 of TMM normalized data (logFC) was used to rank the data from top upregulated to top downregulated genes and FDR (0.05) was used to define

significantly differentially expressed genes. RNA-seq data have been deposited in NCBI's Gene Expression Omnibus (GEO) and are accessible through GEO Series accession number GSE70685.

GSEA and iRegulon AHR target prediction. iRegulon²⁶ was used to retrieve the top 100 AHR predicted targets with a minimal occurrence count threshold of 5. The data were analysed using GSEA²⁷ with ranked data as input with parameters set to 2,000 gene-set permutations.

GSEA and StemRegenin 1 (SR1) gene sets. The GEO dataset GSE28359, which contains Affymetrix Human Genome U133 Plus 2.0 Array gene expression data for CD34⁺ cells treated with SR1 at 30 nM, 100 nM, 300 nM and 1,000 nM was used to obtain lists of genes differentially expressed in the treated samples compared to the control ones (0 nM)². Data were background corrected using Robust Multi-Array Average (RMA) and quantile normalized using the *expresso()* function of the affy Bioconductor package (affy_1.38.1, R 3.0.1). Lists of genes were created from the 150 top upregulated and downregulated genes from the SR1-treated samples at each dose compared to the non-treated samples (0 nM). The data were analysed using GSEA with ranked data as input with parameters set to 2,000 gene-set permutations. The normalized enrichment score (NES) and false discovery rate (FDR) were calculated for each comparison.

Differentiation Map of Haematopoiesis (DMAP) population comparisons. The GEO data set GSE24759, which contains Affymetrix GeneChip HT-HG_U133A Early Access Array gene expression data for 38 distinct haematopoietic cell states⁴, was compared to the MSI2 overexpression and knockdown data. GSE24759 data were background corrected using Robust Multi-Array Average (RMA), quantile normalized using the *expresso()* function of the affy Bioconductor package (affy_1.38.1, R 3.0.1), batch corrected using the *ComBat()* function of the sva package (sva_3.6.0) and scaled using the standard score. Bar graphs were created by calculating for significantly differentially expressed genes the number of scaled data that were above (>0) or below (<0) the mean for each population. Percentages indicating for how long the observed value (set of up- or downregulated genes) was better represented in that population than random values were calculated from 1,000 trials.

AHR ChIP-seq comparison with downregulated gene sets. A unique list of genes closest to AHR-bound regions previously identified from TCDD-treated MCF7 ChIP-seq data¹⁴ was used to calculate the overlap with genes showing >1.5-fold downregulation in response to treatment with UM171 (35 nM) or SR1 (500 nM) relative to DMSO-treated samples³ as well as with genes significantly downregulated in MSI2-overexpressing versus control treated samples (FDR < 0.05). The percentage of downregulated genes with AHR-bound regions was then plotted for each gene set. *P* values were generated with Fisher's exact test for comparisons between gene lists.

oPOSSUM analysis for promoter AHR binding sites in downregulated gene sets. AHR transcription factor binding sites in downregulated gene sets were identified with oPOSSUM-3²⁸. Genes showing >1.5-fold downregulation in response to treatment with UM171 (35 nM) or SR1 (500 nM) relative to DMSO-treated samples³ were used along with significantly downregulated genes (FDR < 0.05) with EdgeR-analysed MSI2-overexpressing versus control-treated samples. The three gene lists were uploaded into oPOSSUM-3 and the AHR:ARNT transcription factor binding site profile was used with the matrix score threshold set at 80% to analyse the region 1,500 bp upstream and 1,000 bp downstream of the transcription start site. The percentage of downregulated genes with AHR-binding sites in their promoters was then plotted for each gene set. Fisher's exact test was used to identify significant overrepresentation of AHR-binding sites in gene lists relative to background.

Analysis for human chimaerism. Eight- to 12-week-old male or female NSG mice were sublethally irradiated (315 cGy) one day before intrafemoral injection with transduced cells carried in IMDM 1% FBS at 25 µl per mouse. Injected mice were analysed for human haematopoietic engraftment 12–14 weeks after transplantation or at 3 and 6.5 weeks for STRC experiments. Mouse bones (femurs, tibiae and pelvis) and spleen were removed and bones were crushed with a mortar and pestle then filtered into single-cell suspensions. Bone marrow and spleen cells were blocked with mouse Fc block (BD Biosciences) and human IgG (Sigma) and then stained with fluorochrome-conjugated antibodies specific to human haematopoietic cells. For multilineage engraftment analysis, cells from mice were stained with CD45 (HI30) (Invitrogen), CD33 (P67.6), CD15 (HI98), CD14 (MφP9), CD19 (HIB19), CD235a/GlyA (GA-R2), CD41a (HIP8) and CD34 (581) (BD Biosciences).

HSC and STRC xenotransplantation. For MSI2 knockdown in HSCs, 5.0×10^4 and 2.5×10^4 sorted Lin[−] CD34⁺ CD38[−] cells were used per short-hairpin transduction experiment, leading to transplantation of day zero equivalent cell doses of 10×10^5 and 6.25×10^5 , respectively, per mouse. For STRC LDA transplantation experiments, 10^5 sorted CD34⁺CD38⁺ cells were used per control or MSI2-overexpressing transduction. After assessing levels of gene transfer, day zero equivalent GFP⁺ cell doses were calculated to perform the LDA. Recipients with greater than 0.1% GFP⁺CD45^{+/−} cells were considered to be repopulated. For STRC experiments that read out extended engraftment at 6.5 weeks, 2×10^5 CD34⁺ CD38⁺ cells were used per overexpressing or control transduction to allow

non-limiting 5×10^4 day zero equivalent cell doses per mouse. For HSC expansion and LDA experiments, CD34⁺CD38[−] cells were sorted and transduced with MSI2-overexpressing or control vectors (50,000 cells per condition) for 3 days and then analysed for gene-transfer levels (% GFP^{+/−}) and primitive cell marker expression (% CD34 and CD133). To ensure that equal numbers of GFP⁺ cells were transplanted into both control and MSI2-overexpressing recipient mice, we added identically cultured GFP[−] cells to the MSI2 culture to match the % GFP⁺ of the control culture (necessary owing to the differing efficiency of transduction). The adjusted MSI2-overexpressing culture was recounted and aliquoted (63,000 cells) to match the output of half of the control culture. Three day 0 equivalent GFP⁺ cell doses (1,000, 300 and 62 cells) were then transplanted per mouse to perform the D3 primary LDA. A second aliquot of the adjusted MSI2-overexpressing culture was then taken and put into culture in parallel with the remaining half of the control culture to perform another LDA after 7 days of growth (10 days total growth, D10 primary LDA). Altogether, four cell doses were transplanted; when converted back to day 0 equivalents these equalled approximately 1,000, 250, 100, and 20 GFP⁺ cells per mouse, respectively. Pooled bone marrow from six engrafted primary mice that received D10 cultured control or MSI2-overexpressing cells (from the two highest doses transplanted) was aliquoted into five cell doses of 15 million, 10 million, 6 million, 2 million and 1 million cells. The numbers of GFP⁺ cells within primary mice was estimated from nucleated cell counts obtained from NSG femurs, tibiae and pelvises and from Colvin *et al.*²⁹. The actual numbers of GFP⁺ cells used for determining numbers of GFP⁺ HSCs and the number of mice transplanted for all LDA experiments is shown in Supplementary Tables 3–5. The cut-off for HSC engraftment was a demonstration of multilineage reconstitution that was set at bone marrow having >0.1% GFP⁺ CD33⁺ and >0.1% GFP⁺ CD19⁺ cells. HSC and STRC frequency was assessed using ELDA software³⁰. For all mouse transplantation experiments, mice were age- (6–12 week) and sex-matched. All transplanted mice were included for analysis unless mice died from radiation sickness before the experimental endpoint. No randomization or blinding was performed for animal experiments. Approximately 3–6 mice were used per cell dose for each cord blood transduction and transplantation experiment.

UV CLIP-seq library preparation. CLIP-seq was performed as previously described¹⁵. Briefly, 25 million NB4 cells (a transformed human cell line of haematopoietic origin) were washed in PBS and UV-cross-linked at 400 mJ cm^{−2} on ice. Cells were pelleted, lysed in wash buffer (PBS, 0.1% SDS, 0.5% Na-deoxycholate, 0.5% NP-40) and DNase-treated, and supernatants from lysates were collected for immunoprecipitation. MSI2 was immunoprecipitated overnight using 5 µg of anti-MSI2 antibody (EP1305Y, Abcam) and Protein A Dynabeads (Invitrogen). Beads containing immunoprecipitated RNA were washed twice with wash buffer, high-salt wash buffer (5× PBS, 0.1% SDS, 0.5% Na-Deoxycholate, 0.5% NP-40), and PNK buffer (50 mM Tris-Cl pH 7.4, 10 mM MgCl₂, 0.5% NP-40). Samples were then treated with 0.2 U MNase for 5 min at 37° with shaking to trim immunoprecipitated RNA. MNase inactivation was then carried out with PNK + EGTA buffer (50 mM Tris-Cl pH 7.4, 20 mM EGTA, 0.5% NP-40). The sample was dephosphorylated using alkaline phosphatase (CIP, NEB) at 37° for 10 min followed by washing with PNK+EGTA, PNK buffer, and then 0.1 mg ml^{−1} BSA in nuclease-free water. 3'RNA linker ligation was performed at 16° overnight with the following adaptor: 5'P-UGGAUUCUCGGUGCCAAGG-puromycin. Samples were then washed with PNK buffer, radiolabelled using P32-γ-ATP (Perkin Elmer), run on a 4–12% Bis-Tris gel and then transferred to a nitrocellulose membrane. The nitrocellulose membrane was developed via autoradiography and RNA–protein complexes 15–20 kDa above the molecular weight of MSI2 were extracted with proteinase K followed by RNA extraction with acid phenol-chloroform. A 5'RNA linker (5'HO-GUUCAGAGUUCUACAGUCCGACGAUC-OH) was ligated to the extracted RNA using T4 RNA ligase (Fermentas) for 2 h and the RNA was again purified using acid phenol-chloroform. Adaptor ligated RNA was re-suspended in nuclease-free water and reverse transcribed using Superscript III reverse transcriptase (Invitrogen). Twenty cycles of PCR were performed using NEB Phusion Polymerase using a 3'PCR primer that contained a unique Illumina barcode sequence. PCR products were run on an 8% TBE gel. Products ranging between 150 and 200 bp were extracted using the QIAquick gel extraction kit (Qiagen) and re-suspended in nuclease-free water. Two separate libraries were prepared and sent for single-end 50-bp Illumina sequencing at the Institute for Genomic Medicine at the University of California, San Diego. 47,098,127 reads from the first library passed quality filtering, of which 73.83% mapped uniquely to the human genome. 57,970,220 reads from the second library passed quality filtering, of which 69.53% mapped uniquely to the human genome. CLIP-data reproducibility was verified through high correlation between gene RPKMs and statistically significant overlaps in the clusters and genes within replicates. CLIP-seq data have been deposited in NCBI's GEO and are accessible through GEO Series accession number GSE69583.

CLIP-seq mapping and cluster identification. Before sequence alignment of CLIP-seq reads to the human genome was performed, sequencing reads from

libraries were trimmed of polyA tails, adapters, and low quality ends using Cutadapt with parameters—match-read-wildcards—times 2 -e 0 -O 5—quality-cutoff 6 -m 18 -b TCGTATGCCGTCTTCTGCTTG -b ATCTCGTATGCCGTCTTCTGCTTG -b CGACAGGTTTCAGAGTTCTACAGTCCGACGATC -b TGGAATTCTC GGGTGCCAAAGG -b AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA AAAAAAAAAAAAAAAAAA-b TTTTTTTTTTTTTTTTTTTTTTTTTTTTTT TTTTTTTTTTTTTTTTTTTTTT. Reads were then mapped against a database of repetitive elements derived from RepBase (version 18.05). Bowtie (version 1.0.0) with parameters -S -q -p 16 -e 100 -l 20 was used to align reads against an index generated from Repbase sequences³¹. Reads not mapped to Repbase sequences were aligned to the hg19 human genome (UCSC assembly) using STAR (version 2.3.0e)³² with parameters—outSAMunmapped Within—outFilterMultimapNmax 1—outFilterMultimapScoreRange 1. To identify clusters in the genome of significantly enriched CLIP-seq reads, reads that were PCR replicates were removed from each CLIP-seq library using a custom script of the same method as in ref. 33; otherwise, reads were kept at each nucleotide position when more than one read's 5'-end was mapped. Clusters were then assigned using the CLIPper software with parameters—bonferroni-superlocal-threshold-³⁴. The ranked list of significant targets was calculated assuming a Poisson distribution, where the observed value is the number of reads in the cluster, and the background is the number of reads across the entire transcript and or across a window of 1000 bp ± the predicted cluster.

Gene annotations for CLIP-seq. Transcriptomic regions and gene classes were defined using annotations found in genecode v17. Depending on the analysis, clusters were associated by the Gencode-annotated 5'UTR, 3'UTR, CDS or intronic regions. If a cluster overlapped multiple regions, or a single part of a transcript was annotated as multiple regions, clusters were iteratively assigned first as CDS, then 3'UTR, 5'UTR and finally as proximal (<500 bases from an exon) or distal (>500 bases from an exon) introns. Overlapping peaks were calculated using bedtools and pybedtools^{35,36}.

Gene ontology analysis for CLIP-seq. Significantly enriched gene ontology (GO) terms were identified using a hypergeometric test that compared the number of genes that were MSI2 targets in each GO term to genes expressed in each GO term as the proper background. Expressed genes were identified using the control samples in SRA study SRP012062. Mapping was performed identically to CLIP-seq mapping, without peak calling and changing the STAR parameter outFilterMultimapNmax to 10. Counts were calculated with featureCounts³⁷ and RPKMs were then computed. Only genes with a mean RPKM > 1 between the two samples were used in the background expressed set.

De novo motif and conservation analysis for CLIP-seq. Randomly located clusters within the same genic regions as predicted MSI2 clusters were used to calculate a background distribution for motif and conservation analyses. Motif analysis was performed using the HOMER algorithm as in ref. 34. For evolutionary sequence conservation analysis, the mean (mammalian) phastCons score for each cluster was used.

Immunofluorescence. CD34⁺ cells (>5 × 10⁴) were transduced with an MSI2-overexpression or control lentivirus. Three days later, GFP⁺ cells were sorted and then put back in to StemSpan medium containing growth factors IL-6 (20 ng ml⁻¹), SCF (100 ng ml⁻¹), FLT3-L (100 ng ml⁻¹) and TPO (20 ng ml⁻¹). A minimum of 10,000 cells were used for immunostaining at culture days 3 and 7 after GFP sorting. Cells were fixed in 2% PFA for 10 min, washed with PBS and then cytospun on to glass slides. Cytospun cells were then permeabilized (PBS, 0.2% Triton X-100) for 20 min, blocked (PBS, 0.1% saponin, 10% donkey serum) for 30 min and stained with primary antibodies (CYP1B1 (EPR14972, Abcam); HSP90 (68/hsp90, BD Biosciences)) in PBS with 10% donkey serum for 1 h. Detection with secondary antibody was performed in PBS 10% donkey serum with Alexa-647 donkey anti-rabbit antibody or Alexa-647 donkey anti-mouse antibodies for 45 min. Slides were mounted with Prolong Gold Antifade containing DAPI (Invitrogen). Several images (200–1,000 cells total) were captured per slide at 20× magnification using an Operetta HCS Reader (Perkin Elmer) with epifluorescence illumination and standard filter sets. Columbus software (Perkin Elmer) was used to automate the identification of nuclei and cytoplasm boundaries in order to quantify mean cell fluorescence.

Luciferase reporter gene assay. A 271-bp region of the CYP1B1 3'UTR that flanked CLIP-seq-identified MSI2-binding sites was cloned from human HEK293FT genomic DNA using the forward primer GTGACACAACGTGTGATTAAAGG and reverse primer TGATTTTATTTATTTTGGT AATGGTG and placed downstream of renilla luciferase in the dual-luciferase reporter vector pGL4 (Promega). A 271-bp geneblock (IDT) with 6 TAG > TCC mutations was cloned in to pGL4 using XbaI and NotI. The HSP90 3'UTR was amplified from HEK293FT genomic DNA with the forward primer TCTCTGGCTGAGGGATGACT and reverse primer TTTTAAGGCCAAGGAATTAAGTGA and cloned into pGL4. A geneblock of the HSP90 3'UTR (IDT) with 14 TAG > TCC mutations was cloned in to pGL4 using SfaI and NotI. Co-transfection of wild-type or mutant luciferase reporter (40 ng) and control or MSI2-overexpressing lentiviral expression vector (100 ng) was performed in the NIH-3T3 cell line, which does not express MSI1 or

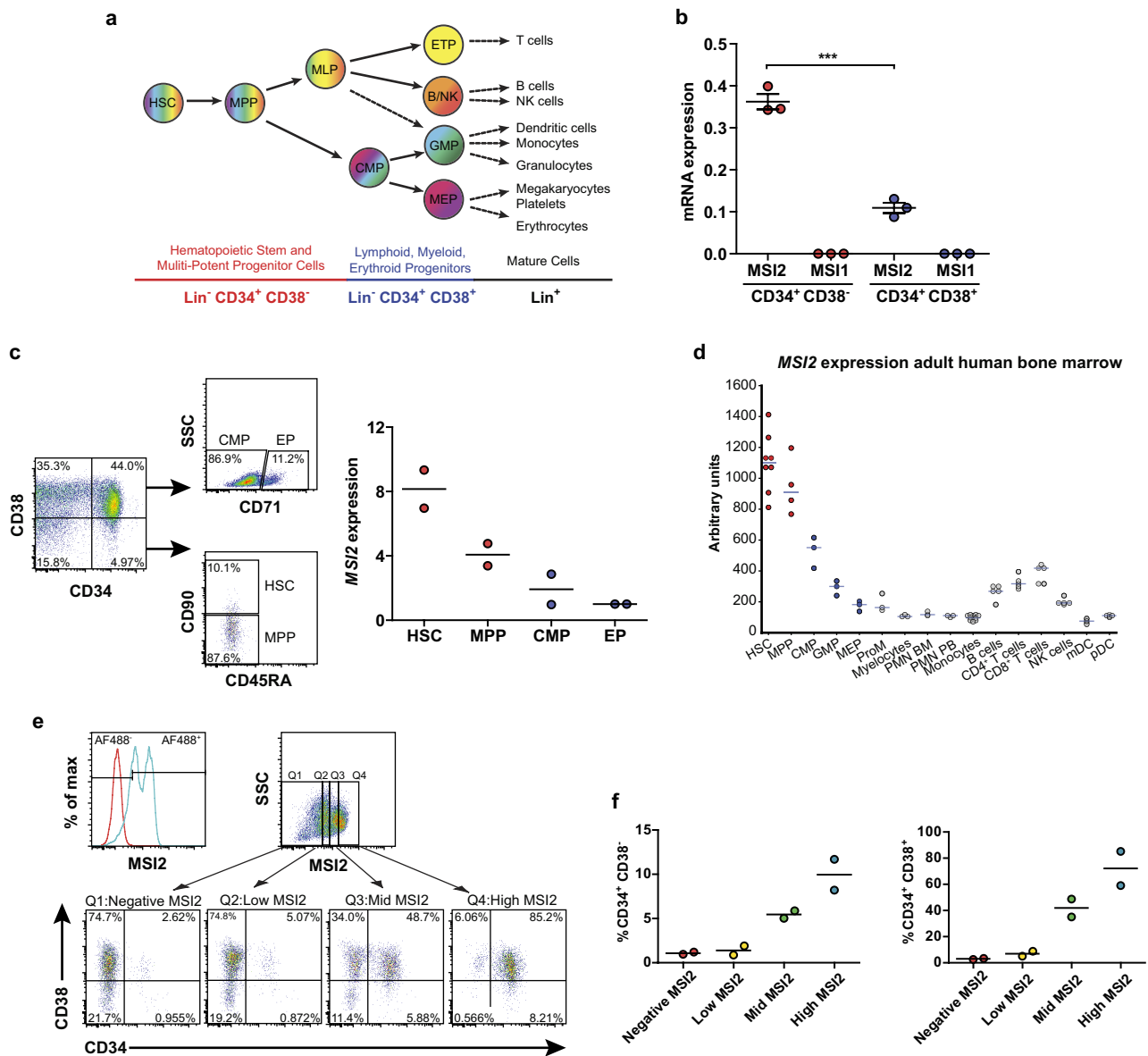
MSI2 (50,000 cells per co-transfection). Reporter activity was measured using the Dual-Luciferase Reporter Assay System (Promega) 36–40 h later.

MSI2-overexpressing suspension cultures with the AHR antagonist SR1 and agonist FICZ. For MSI2-overexpressing cultures with the AHR antagonist SR1, Lin⁻ CD34⁺ cells were transduced with MSI2-overexpression or control lentivirus in medium supplemented with SR1 (750 nM; Abcam) or DMSO vehicle (0.1%). GFP⁺ cells were isolated (20,000 cells per culture) and allowed to proliferate with or without SR1 for an additional 7 days at which point they were counted and immunophenotyped for CD34 and CD133 expression. For MSI2-overexpressing cultures with the AHR agonist FICZ, Lin⁻ CD34⁺ cells were transduced with MSI2-overexpression or control lentivirus. GFP⁺ cells were isolated (20,000 cells per culture) and allowed to proliferate with FICZ (200 nM; Santa Cruz Biotechnology) or DMSO (0.1%) for an additional 3 days, at which point they were immunophenotyped for CD34 and CD133 expression.

HSPC expansion with (E)-2,3',4,5'-tetramethoxystilbene (TMS). Lin⁻ CD34⁺ cells were cultured for 72 h (lentiviral treated but non-transduced flow-sorted GFP⁻ cells) in StemSpan medium containing growth factors IL-6 (20 ng ml⁻¹), SCF (100 ng ml⁻¹), FLT3-L (100 ng ml⁻¹) and TPO (20 ng ml⁻¹) before the addition of the CYP1B1 inhibitor TMS (Abcam) at a concentration of 10 μM or mock treatment with 0.1% DMSO. Equal numbers of cells (12,000 per condition) were then allowed to proliferate for 7 days at which point they were counted and immunophenotyped for CD34 and CD133 expression.

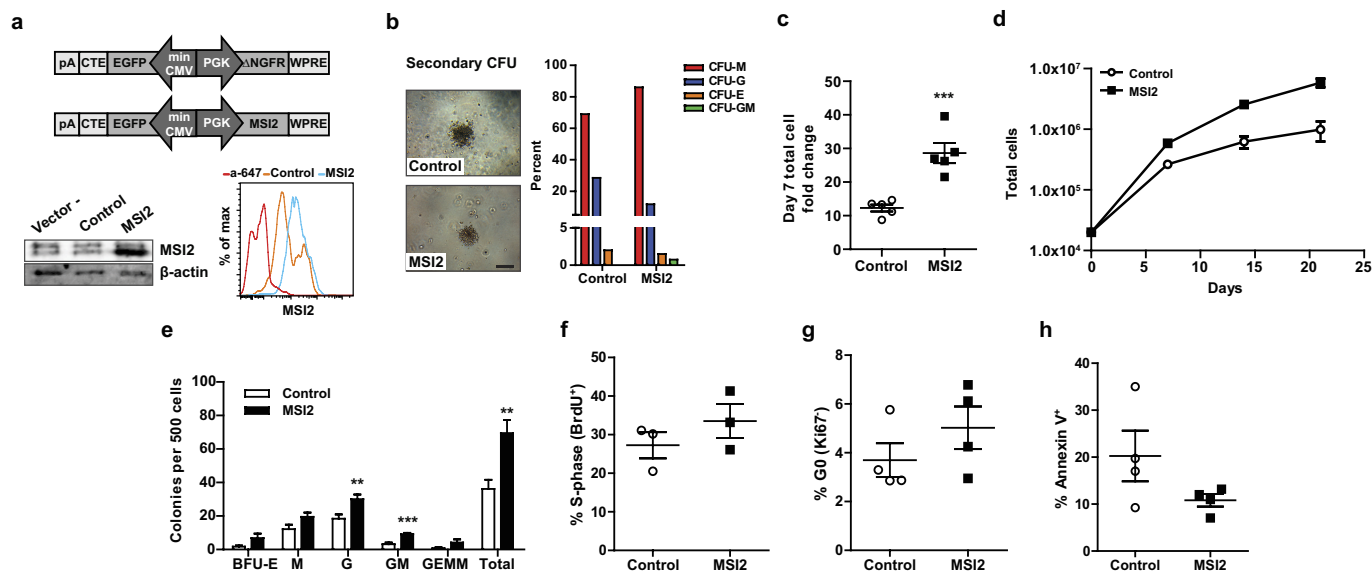
Statistical analysis. Unless stated otherwise (that is, analysis of RNA-seq and CLIP-seq data sets), all statistical analysis was performed using GraphPad Prism (GraphPad Software version 5.0). Unpaired student *t*-tests or Mann-Whitney tests were performed with *P* < 0.05 as the cut-off for statistical significance. No statistical methods were used to predetermine sample size.

- Doulavov, S. *et al.* PLZF is a regulator of homeostatic and cytokine-induced myeloid development. *Genes Dev.* **23**, 2076–2087 (2009).
- Majeti, R., Park, C. Y. & Weissman, I. L. Identification of a hierarchy of multipotent hematopoietic progenitors in human cord blood. *Cell Stem Cell* **1**, 635–645 (2007).
- Bagger, F. O. *et al.* HemaExplorer: a database of mRNA expression profiles in normal and malignant haematopoiesis. *Nucleic Acids Res.* **41**, D1034–D1039 (2013).
- Amendola, M., Venneri, M. A., Biffi, A., Vigna, E. & Naldini, L. Coordinate dual-gene transgenesis by lentiviral vectors carrying synthetic bidirectional promoters. *Nature Biotechnol.* **23**, 108–116 (2005).
- van Galen, P. *et al.* The unfolded protein response governs integrity of the haematopoietic stem-cell pool during stress. *Nature* **510**, 268–272 (2014).
- Lechman, E. R. *et al.* Attenuation of miR-126 activity expands HSC in vivo without exhaustion. *Cell Stem Cell* **11**, 799–811 (2012).
- Carow, C. E., Hangoc, G. & Broxmeyer, H. E. Human multipotential progenitor cells (CFU-GEMM) have extensive replating capacity for secondary CFU-GEMM: an effect enhanced by cord blood plasma. *Blood* **81**, 942–949 (1993).
- Milyavsky, M. *et al.* A distinctive DNA damage response in human hematopoietic stem cells reveals an apoptosis-independent role for p53 in self-renewal. *Cell Stem Cell* **7**, 186–197 (2010).
- Janky, R. *et al.* iRegulon: from a gene list to a gene regulatory network using large motif and track collections. *PLOS Comput. Biol.* **10**, e1003731 (2014).
- Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
- Kwon, A. T., Arenillas, D. J., Worsley Hunt, R. & Wasserman, W. W. oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets. *G3* **2**, 987–1002 (2012).
- Colvin, G. A. *et al.* Murine marrow cellularity and the concept of stem cell competition: geographic and quantitative determinants in stem cell biology. *Leukemia* **18**, 575–583 (2004).
- Hu, Y. & Smyth, G. K. ELDA: extreme limiting dilution analysis for comparing depleted and enriched populations in stem cell and other assays. *J. Immunol. Methods* **347**, 70–78 (2009).
- Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
- Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
- Darnell, R. CLIP (cross-linking and immunoprecipitation) identification of RNAs bound by a specific protein. *Cold Spring Harb. Protoc.* **2012**, 1146–1160 (2012).
- Lovci, M. T. *et al.* Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nature Struct. Mol. Biol.* **20**, 1434–1442 (2013).
- Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
- Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**, 3423–3424 (2011).
- Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).



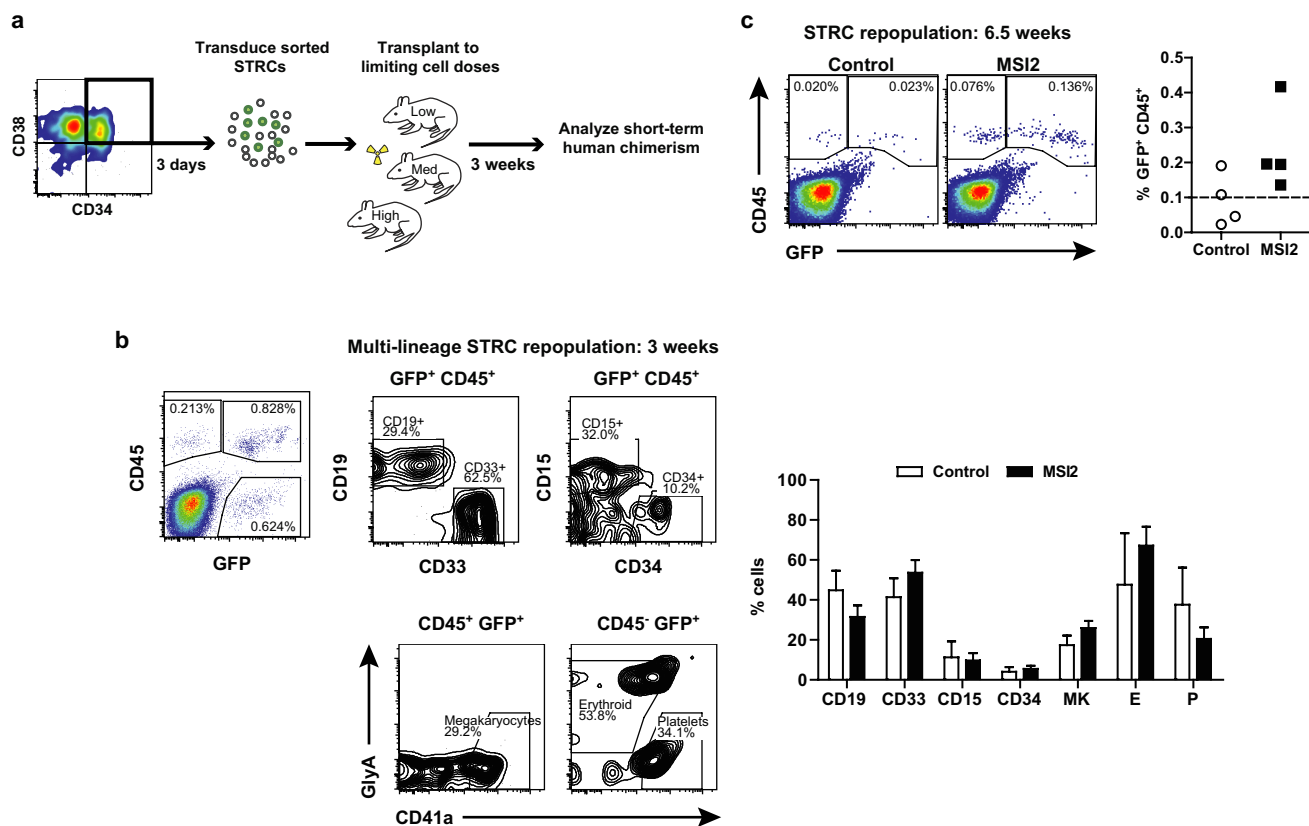
Extended Data Figure 1 | MSI2 is highly expressed in human haematopoietic stem and progenitor cell populations. **a**, Schematic of the human haematopoietic hierarchy showing key primitive cell populations and simplified surface marker expression. **b**, qRT-PCR analysis of *MSI1* and *MSI2* expression in Lin^- cord blood (CB) cell populations ($n = 3$ independent Lin^- CB samples). **c**, Gating strategy used to sort sub-fractions of Lin^- CB HSPCs for *MSI2* qRT-PCR expression analysis ($n = 2$ independent pooled Lin^- CB samples). **d**, *MSI2* expression across the human haematopoietic hierarchy. Each circle represents an

independent gene expression data set curated by HemaExplorer. **e**, Intracellular flow cytometry analysis of *MSI2* protein levels in Lin^- CB. Histograms show background staining with secondary antibody (red) and positive staining with anti-*MSI2* antibody plus secondary in Lin^- CB (blue). *MSI2* fluorescence intensity was divided into quartiles of negative (Q1), low (Q2), mid (Q3) and high (Q4) level expression. **f**, Plots show cell percentage within each quartile from **e** that are $\text{CD34}^+ \text{CD38}^-$ (left) and $\text{CD34}^+ \text{CD38}^+$ (right) ($n = 2$ independent Lin^- CB samples). All data presented as mean \pm s.e.m. Unpaired *t*-test, * $P < 0.05$; *** $P < 0.001$.



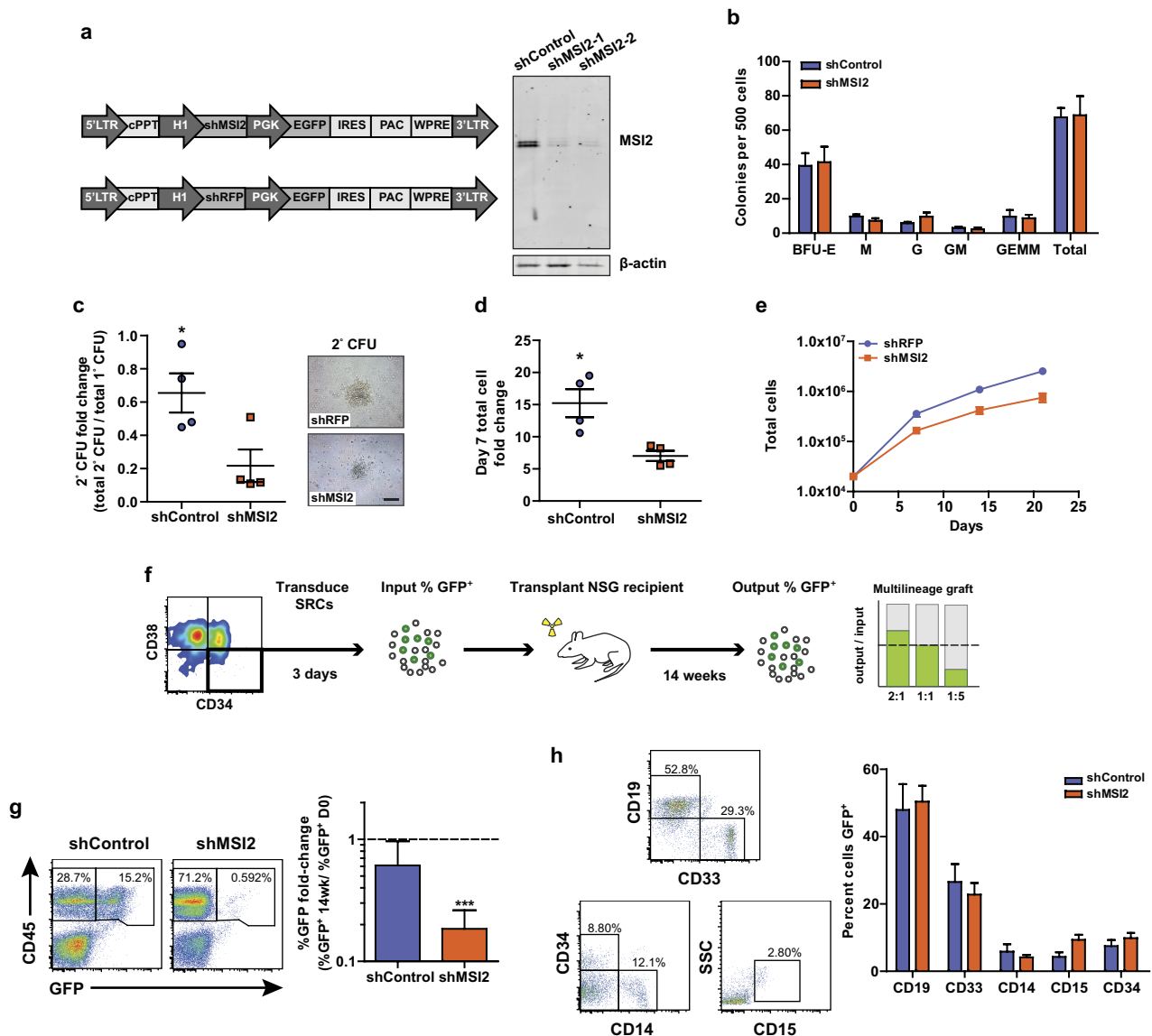
Extended Data Figure 2 | MSI2 overexpression enhances *in vitro* culture of primitive CB cells. **a**, Top: schematic of bi-directional promoter lentivirus used to overexpress MSI2. Bottom: western blot and histogram showing intracellular flow validation of enforced MSI2 expression in 293FT cells (left) and Lin⁻ CB (right), respectively. **b**, Representative images of secondary CFU made from replated control and MSI2-overexpressing (MSI2) CFU-GEMMs and types of colonies made. Scale bar, 200 μm. **c**, Fold change in Lin⁻ CB transduced cell number after 7 days in culture following transduction ($n = 5$ experiments). **d**, Growth curve

over 21 days of transduced Lin⁻ CB cells ($n = 4$ experiments). **e**, Colony output of transduced Lin⁻ CB from day 7 cultures ($n = 8$ cultures from 4 experiments). **f**, BrdU cell cycle analysis of transduced Lin⁻ CB cells from day 10 cultures ($n = 3$ experiments). **g**, Ki67 cell cycle analysis of transduced Lin⁻ CB cells from day 4 cultures ($n = 4$ experiments). **h**, Apoptotic and dead cells in day 7 cultures of transduced Lin⁻ CB by Annexin V staining ($n = 3$ experiments). Western blot source data are available in Supplementary Fig. 1. All data presented as mean \pm s.e.m. Unpaired t -test, ** $P < 0.01$; *** $P < 0.001$.



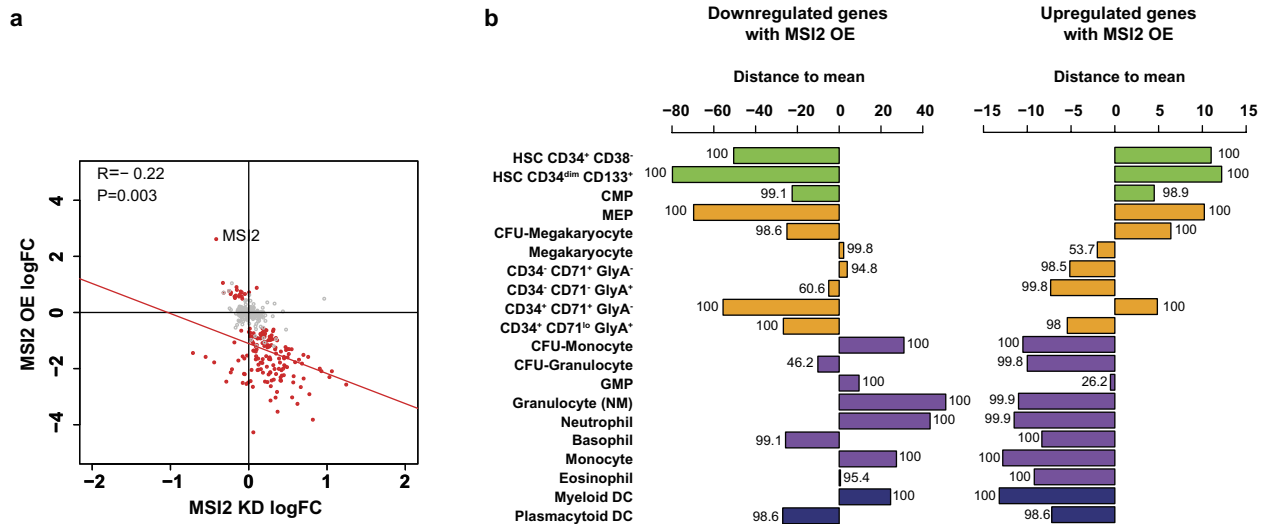
Extended Data Figure 3 | MSI2 overexpression does not affect STRC lineage output and extends STRC-mediated engraftment. **a**, Schematic of STRC LDA experimental setup. **b**, Left: gating strategy to identify engrafted GFP⁺ CD45⁺ progenitor and myelo-lympho lineage-positive cell types or GFP⁺ CD45⁻ erythroid cells and platelets. Right: summary of lineage output in the injected femur 3 weeks after transplantation

($n = 4$ mice for control and $n = 18$ mice for MSI2 overexpressing cells). MK, megakaryocyte; E, erythroid cells; P, platelets. **c**, Representative flow plots and summary of transduced STRC read out for engraftment with human CD45⁺ cells at 6.5 weeks post-transplant ($n = 4$ mice per condition). All data presented as mean \pm s.e.m.



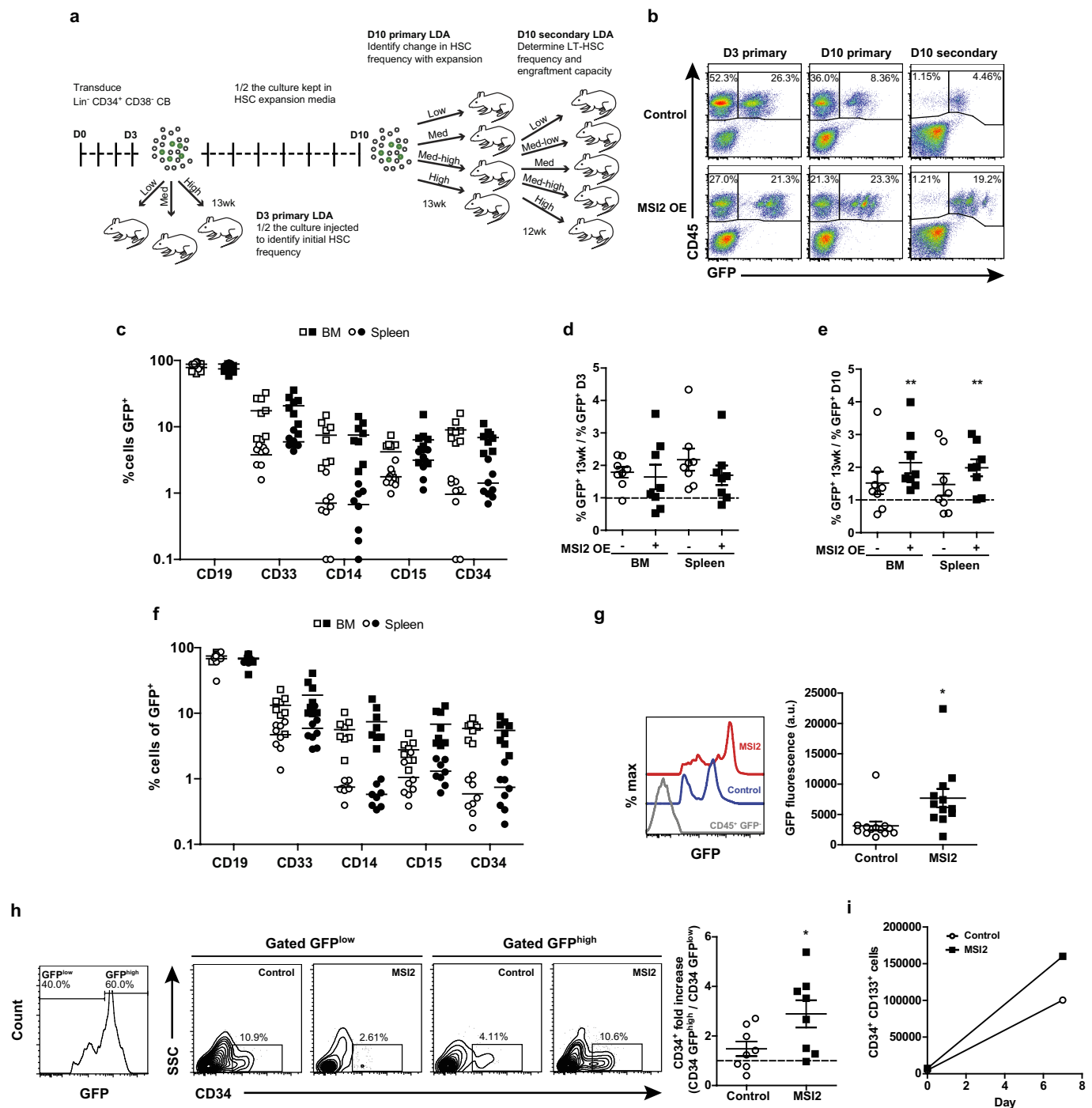
Extended Data Figure 4 | MSI2 knockdown impairs secondary CFU replating potential and HSC engraftment capacity. **a**, Left: schematic of MSI2- and control RFP-targeted shRNA lentiviruses. Right: confirmation of MSI2 protein knockdown (both isoforms that can be detected by western blot) in transduced NB4 cells. **b**, CFU production by shMSI2- and shControl-transduced Lin⁻ CB (*n* = 8 cultures from 4 experiments). **c**, Secondary CFU output from shMSI2-transduced Lin⁻ CB and images of representative secondary CFUs (scale bar, 200 μ m; performed on *n* = 4 cultures from 2 experiments). **d**, Fold change in transduced cell number after 7 days in culture (*n* = 4 experiments). **e**, Growth curves of cultures initiated with transduced Lin⁻ CB cells (*n* = 4 experiments). **f**, Experimental design to read out changes in HSC capacity with MSI2

knockdown. **g**, Left: representative flow analysis of transduced CD34⁺ CD38⁻ derived human chimaerism in NSG mouse bone marrow. Right: ratio of the percentage of GFP⁺ cells in the CD45⁺ population post-transplant to the initial pre-transplant GFP⁺ cell percentage. Dotted line indicates that the proportion of GFP⁺ cells is unchanged relative to input. One sample *t*-test, no change = 1; *n* = 6 mice receiving shControl and *n* = 8 mice receiving shMSI2-transduced cells pooled from two experiments. **h**, Representative flow plots and summary of multilineage engraftment with shControl and shMSI2 cells (gated on GFP⁺ cells). Western blot source data are shown in Supplementary Fig. 1. Data presented as mean \pm s.e.m. Unpaired *t*-test, **P* < 0.05; ****P* < 0.001.



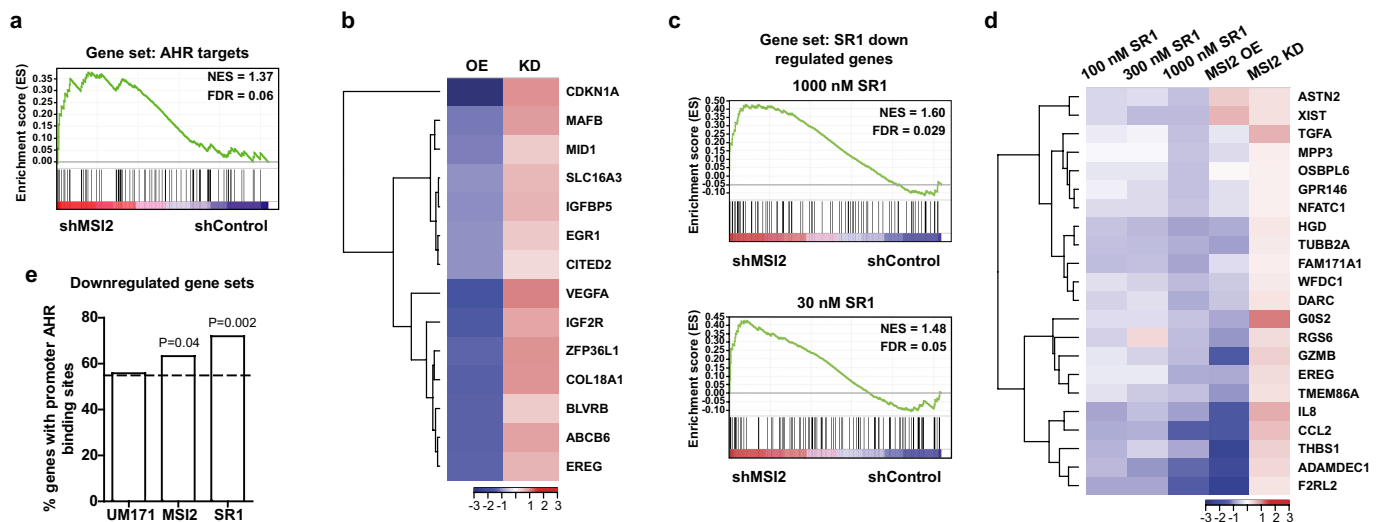
Extended Data Figure 5 | MSI2 overexpression confers an HSC gene expression signature. **a**, Genes that are upregulated (21 genes, logFC > 0) or downregulated (156 genes, logFC < 0) in MSI2-overexpressing (OE) cells relative to control cells with FDR < 0.05 were compared to expression data from MSI2 knockdown cells normalized to shControl expression data. Red circles represent 177 genes that were significantly differentially expressed in MSI2-overexpressing cells. Gray outlined circles represent random genes (equal number of grey circles and red circles). Only genes

that were significantly up- or downregulated in MSI2-overexpressing cells showed anti-correlation with MSI2 knockdown cells. **b**, Genes that were differentially expressed between MSI2-overexpressing and control cells (FDR < 0.05) compared to DMAP populations. Numbers beside each bar indicate the percentage of time for which the observed value (set of up- or downregulated genes) was better represented in that population than random values (equal number of randomly selected genes based on 1,000 trials).



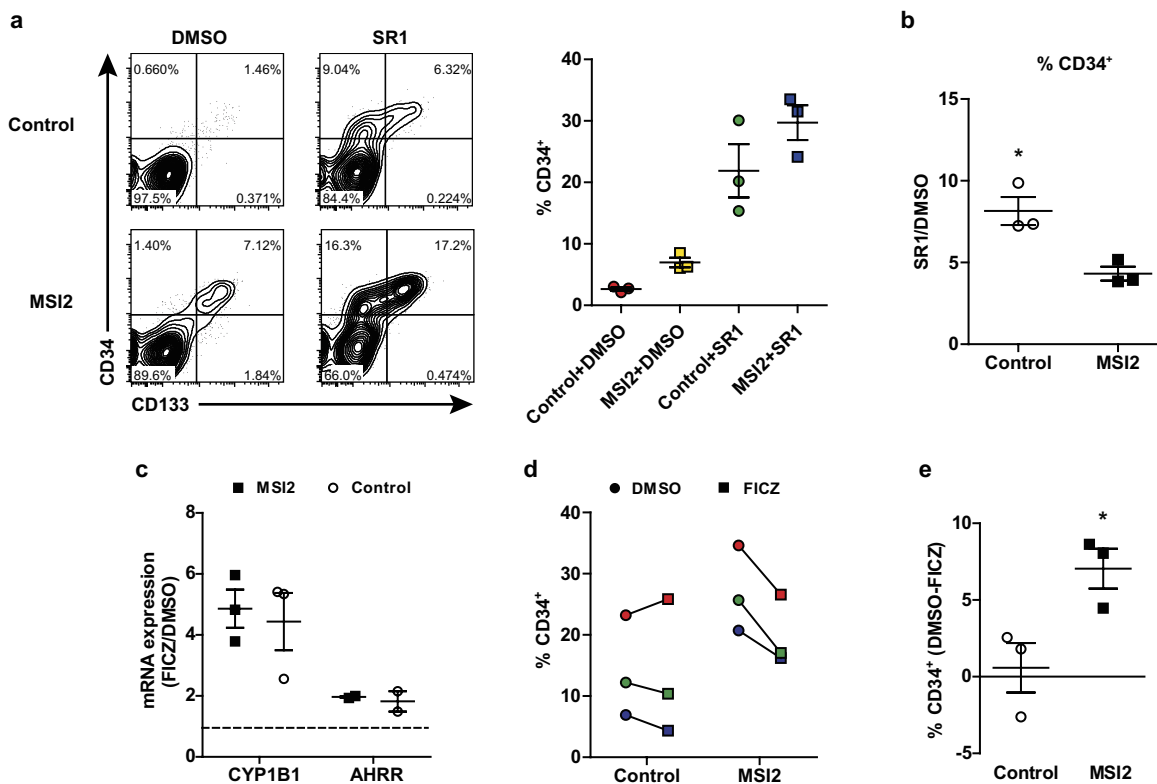
Extended Data Figure 6 | MSI2 overexpression enhances HSC activity after *ex vivo* culture. **a**, Experimental procedure for measuring changes in HSC engraftment capacity and frequency with *ex vivo* culture. **b**, Representative flow plots of CD45⁺ GFP⁺ reconstitution from mice receiving the highest cell dose transplanted per time point. **c**, Multilineage engraftment of mice injected with D3 cultures. **d**, Proportion of the human CD45⁺ graft containing GFP⁺ cells from mice receiving the two highest doses of D3 primary grafts relative to pre-transplant levels of GFP⁺ cells before expansion ($n = 8$ mice for each dose). **e**, Proportion of the human CD45⁺ graft containing GFP⁺ cells from mice receiving the two highest doses of D10 primary grafts relative to pre-transplant levels of

GFP⁺ cells after expansion ($n = 8$ mice for each dose, one-sample *t*-test, no change = 1). **f**, Multilineage engraftment of mice injected with D10 cultures. **g**, GFP mean fluorescence intensity (MFI) in D10 primary cell-engrafted mice. Data are from mice transplanted with the highest three doses; $n = 11$ control and 13 MSI2-overexpressing cell-engrafted mice. **h**, CD34 expression in GFP^{high} (top 60%) relative to GFP^{low} (bottom 40%) gated cells (set per mouse) from engrafted recipients in **e**. **i**, Number of transduced phenotyped HSCs after 7 days of culture from HSC expansion experiment described in **a**. Symbols represent individual mice and shaded symbols represent mice grafted with MSI2-overexpressing cells. All data presented as mean \pm s.e.m. Unpaired *t*-test, * $P < 0.05$.



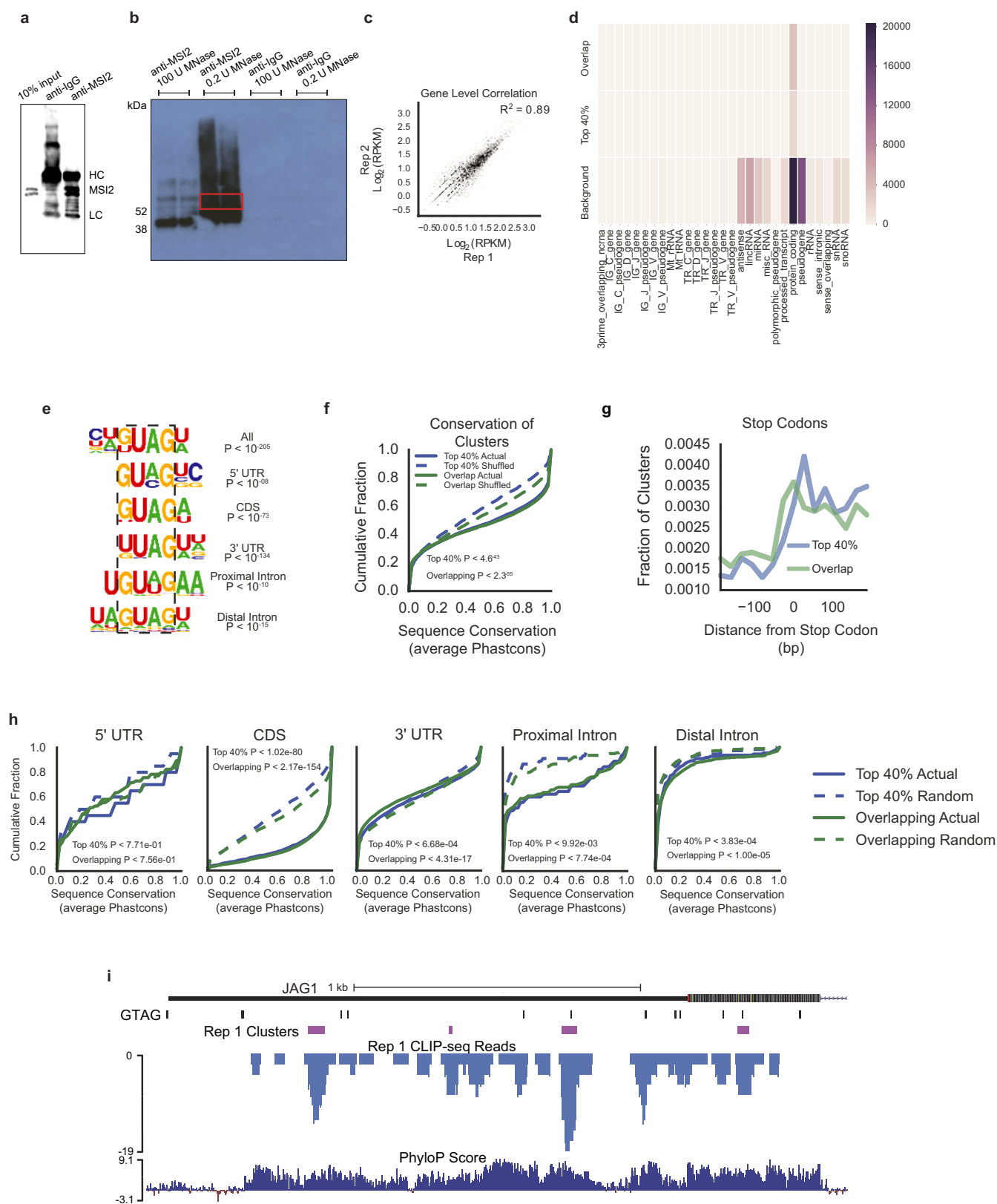
Extended Data Figure 7 | Predicted AHR targets and genes downregulated by SR1 or MSI2 overexpression are upregulated by MSI2 knockdown. **a**, Predicted AHR targets were identified with the iRegulon tool and compared with MSI2 knockdown normalized to shControl-upregulated gene signature by GSEA. **b**, log fold-change of MSI2-overexpression and knockdown shared leading edge AHR target genes from GSEA. **c**, GSEA comparing gene sets downregulated by SR1 high and low dose with the MSI2 knockdown upregulated gene signature.

d, Heatmap and log fold-change of shared leading edge genes identified by GSEA from MSI2 overexpression, MSI2 knockdown and SR1 at varying doses. **e**, The percentage of downregulated genes in UM171-treated, SR1-treated and MSI2-overexpressing cells containing at least one AHR-binding site within 1,500 bp upstream or downstream of the transcription start site. Dotted line indicates the background percentage of genes with AHR-binding sites. *P* values were generated relative to background with Fisher's exact test.



Extended Data Figure 8 | AHR antagonism with SR1 has redundant effects with MSI2 overexpression, and AHR activation with MSI2 overexpression results in a loss of HSPCs. **a**, Representative flow plots and summary of CD34 expression in MSI2-overexpressing and control transduced CD34⁺ CB cells grown for 10 days in medium containing SR1 or DMSO vehicle ($n = 3$ experiments). **b**, Fold change in CD34 expression from results shown in **a**. **c**, Fold increase in *CYP1B1* and *AHRR* transcript

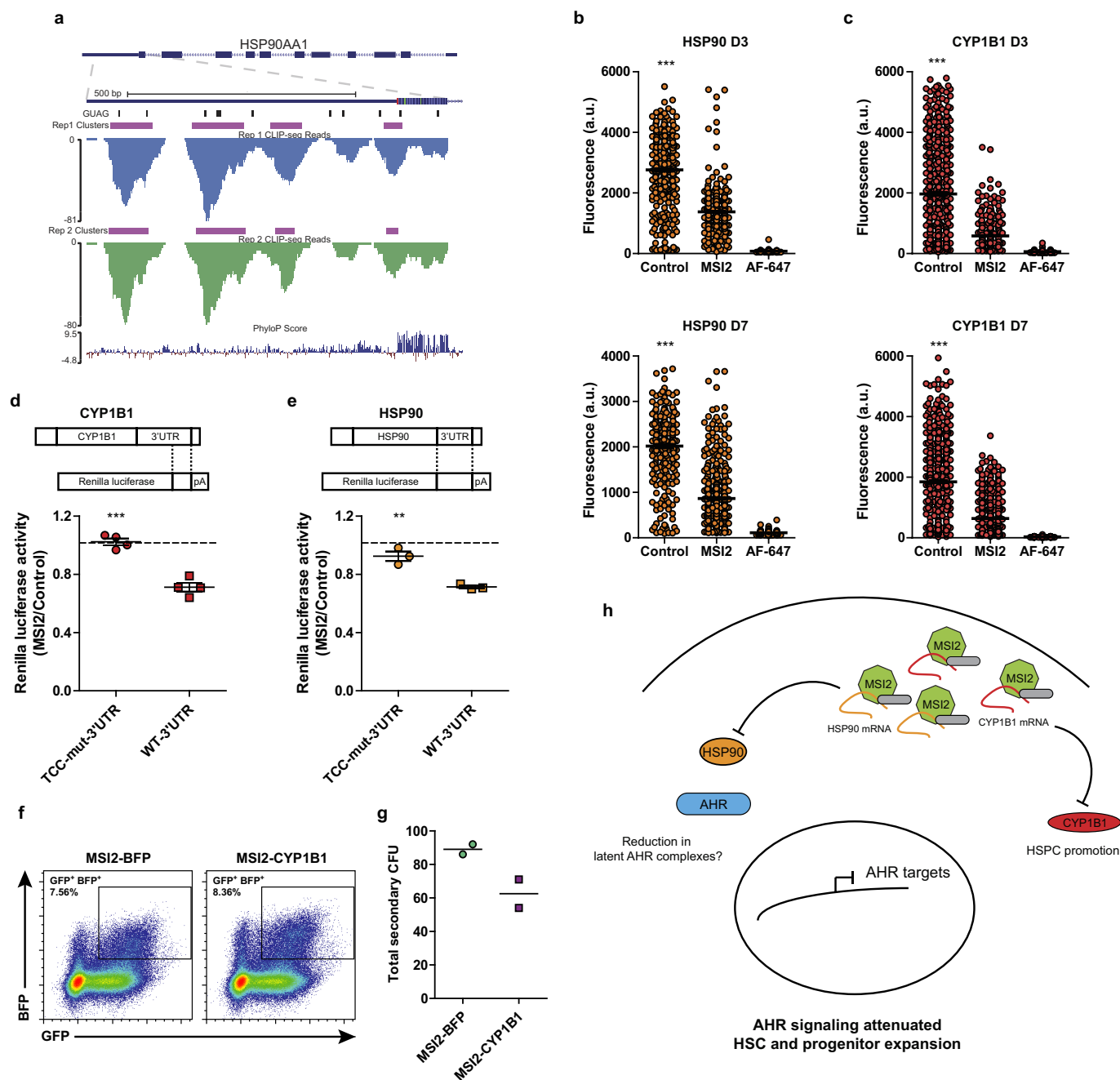
levels after FICZ treatment in transduced cultures ($n = 3$ experiments). **d**, Transduced CD34⁺ CB cells grown for 3 days in medium supplemented with FICZ and the corresponding change in CD34 expression. Each coloured pair (DMSO and FICZ) represents a matched CB sample ($n = 3$ experiments). **e**, Differences in culture CD34 levels from **d**. All data presented as mean \pm s.e.m. Unpaired *t*-test, $*P < 0.05$.



Extended Data Figure 9 | See next page for caption.

Extended Data Figure 9 | MSI2 preferentially binds mature mRNA within the 3'UTR. **a**, Validation of the capacity of the anti-MSI2 antibody to immunoprecipitate MSI2 compared to IgG control pulldowns (heavy chain, HC; light chain, LC). **b**, Autoradiogram showing anti-MSI2 immunoprecipitated, MNase digested and radiolabelled RNA isolated for CLIP library construction and sequencing (red box). Low levels of MNase show a smearing pattern extending upwards from the modal weight of MSI2. **c**, Scatter plot of total number of uniquely mapped CLIP-seq reads for each gene, comparing both replicates. **d**, Heatmap indicating the number of different classes of Gencode-annotated genes that contain at least one predicted MSI-binding site. **e**, Consensus motifs within MSI2 clusters in the different genic regions. *P* values for the most statistically significant enriched motif are presented for all overlapping clusters between replicates. **f**, Cumulative distribution function of mean conservation score (Phastcons) of MSI2 clusters, compared to a shuffled background control, computed for all overlapping clusters and the top 40% of overlapping clusters. *P* values were obtained

by a Kolmogorov–Smirnov two-tailed test comparing the distributions from actual and shuffled locations. **g**, Number of clusters within 200 bases of the annotated stop codon in known mRNA transcripts for all overlapping clusters between replicates and the top 40% of overlapping clusters. **h**, Cumulative distribution function of mean conservation score (Phastcons) of MSI2 clusters, compared to a shuffled background control, computed for overlapping clusters between the replicates and the top 40% of overlapping clusters found in different genic regions. Similarity between the 3'UTR conservation for the top 40% and the shuffled background control is probably due to MSI2 sites being small and not needing structural contexts for conservation. *P* values were obtained by a Kolmogorov–Smirnov two-tailed test comparing the distributions from actual and shuffled locations. **i**, Genome browser views displaying CLIP-seq mapped reads from replicate 1 (blue), predicted clusters (purple), exact matches for the GUAG sequence (black) and mammal conservation scores (PhyloP) in the 3'UTRs for a previously predicted Msi1 target.



Extended Data Figure 10 | MSI2 overexpression represses CYP1B1 and HSP90 3'UTR Renilla Luciferase reporter activity. **a**, CLIP-seq reads (replicate 1 in blue and replicate 2 in green) and clusters (purple) mapped to the 3'UTR of HSP90. Matches to the GUAG motif are shown in black. Mammal PhyloP score listed in last track. **b**, **c**, Representative data of mean per cell fluorescence for HSP90 and CYP1B1 protein in transduced CD34⁺ CB. Protein level in cells during *in vitro* culture was analysed 3 days (D3) and 7 days (D7) after transduction and sorting for GFP. Corresponding secondary-alone antibody staining is shown for each experiment. Each circle represents a cell, and more than 200 cells were analysed per condition. **d**, **e**, Levels of renilla luciferase activity in NIH-3T3 cells co-transfected with control or MSI2 overexpression vectors and the CYP1B1 or HSP90 wild-type or TCC mutant 3'UTR luciferase reporter (dotted

line indicates no change in renilla activity; $n = 4$ CYP1B1 and $n = 3$ HSP90 experiments). **f**, Flow plots of co-transduced CD34⁺ CB cells with MSI2 (GFP) and CYP1B1 (BFP) lentivirus. **g**, GFP⁺ BFP⁺ CFU-GEMMs generated from **f** were replated into secondary CFU assays and the total number of colonies formed was counted. A total of 24 CFU-GEMMs from MSI2-BFP and MSI2-CYP1B1 were replated ($n = 2$ experiments). Data presented as mean \pm s.e.m. Unpaired *t*-test, *** $P < 0.001$, ** $P < 0.01$. **h**, A model for AHR pathway attenuation by MSI2 post-transcriptional processing. MSI2 mediates the post-transcriptional downregulation of HSP90 at the outset of culture and continuously represses the prominent AHR pathway effector CYP1B1 to facilitate HSPC expansion. The resulting MSI2-mediated repression of AHR signalling enforces a self-renewal program and allows HSPC expansion *ex vivo*.

Normalizing the environment recapitulates adult human immune traits in laboratory mice

Lalit K. Beura¹, Sara E. Hamilton², Kevin Bi³, Jason M. Schenkel¹, Oludare A. Odumade^{2†}, Kerry A. Casey^{1†}, Emily A. Thompson¹, Kathryn A. Fraser¹, Pamela C. Rosato¹, Ali Filali-Mouhim⁴, Rafick P. Sekaly⁴, Marc K. Jenkins¹, Vaiva Vezys¹, W. Nicholas Haining³, Stephen C. Jameson² & David Masopust¹

Our current understanding of immunology was largely defined in laboratory mice, partly because they are inbred and genetically homogeneous, can be genetically manipulated, allow kinetic tissue analyses to be carried out from the onset of disease, and permit the use of tractable disease models. Comparably reductionist experiments are neither technically nor ethically possible in humans. However, there is growing concern that laboratory mice do not reflect relevant aspects of the human immune system, which may account for failures to translate disease treatments from bench to bedside^{1–8}. Laboratory mice live in abnormally hygienic specific pathogen free (SPF) barrier facilities. Here we show that standard laboratory mouse husbandry has profound effects on the immune system and that environmental changes produce mice with immune systems closer to those of adult humans. Laboratory mice—like newborn, but not adult, humans—lack effector-differentiated and mucosally distributed memory T cells. These cell populations were present in free-living barn populations of feral mice and pet store mice with diverse microbial experience, and were induced in laboratory mice after co-housing with pet store mice, suggesting that the environment is involved in the induction of these cells. Altering the living conditions of mice profoundly affected the cellular composition of the innate and adaptive immune systems, resulted in global changes in blood cell gene expression to patterns that more closely reflected the immune signatures of adult humans rather than neonates, altered resistance to infection, and influenced T-cell differentiation in response to a *de novo* viral infection. These data highlight the effects of environment on the basal immune state and response to infection and suggest that restoring physiological microbial exposure in laboratory mice could provide a relevant tool for modelling immunological events in free-living organisms, including humans.

Given reported species-specific differences in immune responses^{1–8}, we compared the distribution and differentiation of memory CD8⁺ T cells between mice and humans. CD8⁺ T cells are crucial for adaptive immune control of intracellular infections and cancer, and their distribution and differentiation relate directly to their function. We assessed nonlymphoid distribution by examining available specimens of normal cervical tissue from premenopausal adult women. We found that CD8⁺ T cells were integrated within the mucosa, exhibited a tissue-resident memory T-cell (T_{RM}-cell) phenotype⁹, and comprised ~15,000 of every million cells (Fig. 1a, b and data not shown). These findings are consistent with previous reports that adult (unlike neonatal) human nonlymphoid tissues are abundantly populated with T_{RM} cells^{10–12}. In contrast, CD8⁺ T cells were almost completely absent from cervical sections from adult inbred laboratory mice (C57BL/6 strain) housed under SPF conditions (Fig. 1a, b). We then compared major

CD8⁺ T-cell lineages in the blood of adult humans and laboratory mice, focusing on species-specific markers that define functionally homologous populations of naive, central memory (T_{CM}), and terminally differentiated effector memory CD8⁺ T (T_{EM} or T_{EMRA}) cells (Fig. 1c). Memory CD8⁺ T cells were much scarcer in laboratory mice than in adult humans and were almost entirely comprised of T_{CM} rather than T_{EM} or T_{EMRA} cells. Also unlike humans, laboratory mice lacked CD27^{lo}/granzyme B⁺ effector differentiated memory CD8⁺ T cells, which are thought to respond most immediately to infection^{13,14} (Fig. 1c). Thus, memory CD8⁺ T cells in laboratory mice were scarcer and strikingly different from those in adult humans and,

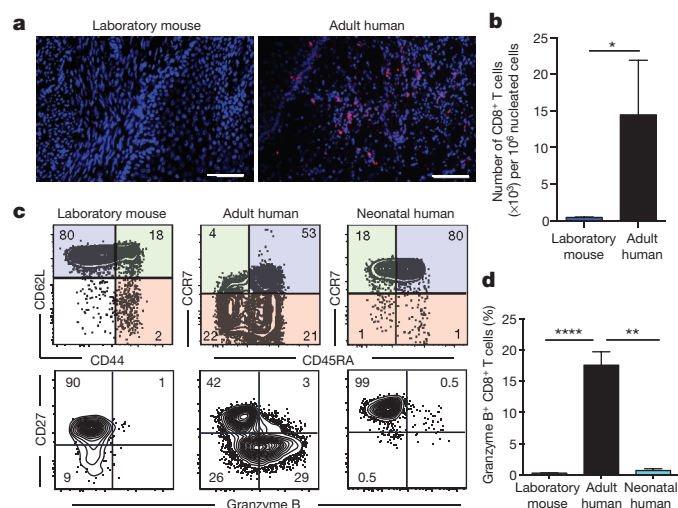


Figure 1 | Laboratory mice, like neonatal but not adult humans, lack differentiated memory CD8⁺ T-cell subsets. **a**, **b**, CD8⁺ T-cell density in cervical tissue from adult laboratory mice ($n = 5$) and humans ($n = 3$). Representative immunofluorescence staining of frozen sections (scale bars, 50 μ m) is shown. Red, CD8 β ; blue, 4',6-diamidino-2-phenylindole (DAPI)-stained nuclei. **c**, CD8⁺ T-cell phenotypes were compared among adult human blood ($n = 13$), adult laboratory mouse blood ($n = 10$), and human cord blood ($n = 8$) in two independent experiments by fluorescence flow cytometry (representative plots shown). Top panels are gated on CD8⁺/CD3⁺ cells and highlight naive (blue), T_{CM} (green) and T_{EM} or T_{EMRA} (red) cells, as defined by conventional lineage markers in each species. Bottom panels are gated on antigen-experienced subsets (green and red quadrants defined above). **d**, Enumeration of granzyme B⁺/CD8⁺ T-cell frequencies in antigen-experienced subsets. Significance was determined using unpaired two-sided Mann–Whitney *U*-test (**b**) or Kruskal–Wallis (analysis of variance, ANOVA) test (**d**). * $P < 0.05$, ** $P < 0.01$, **** $P < 0.0001$; error bars indicate mean \pm s.e.m.

¹Center for Immunology, Department of Microbiology and Immunology, University of Minnesota, Minneapolis, Minnesota 55414, USA. ²Center for Immunology, Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, Minnesota 55414, USA. ³Department of Pediatric Oncology, Dana-Farber Cancer Institute, and Pediatric Hematology and Oncology, Children's Hospital, Boston, Massachusetts 02115, USA. ⁴Department of Pathology, Case Western Reserve University, Cleveland, Ohio 44106, USA. [†]Present addresses: Department of Pediatrics, University of California San Diego, Rady Children's Hospital, San Diego, California 92123, USA (O.A.O.); Department of Respiratory, Inflammation and Autoimmunity, MedImmune LLC, Gaithersburg, Maryland 20878, USA (K.A.C.).

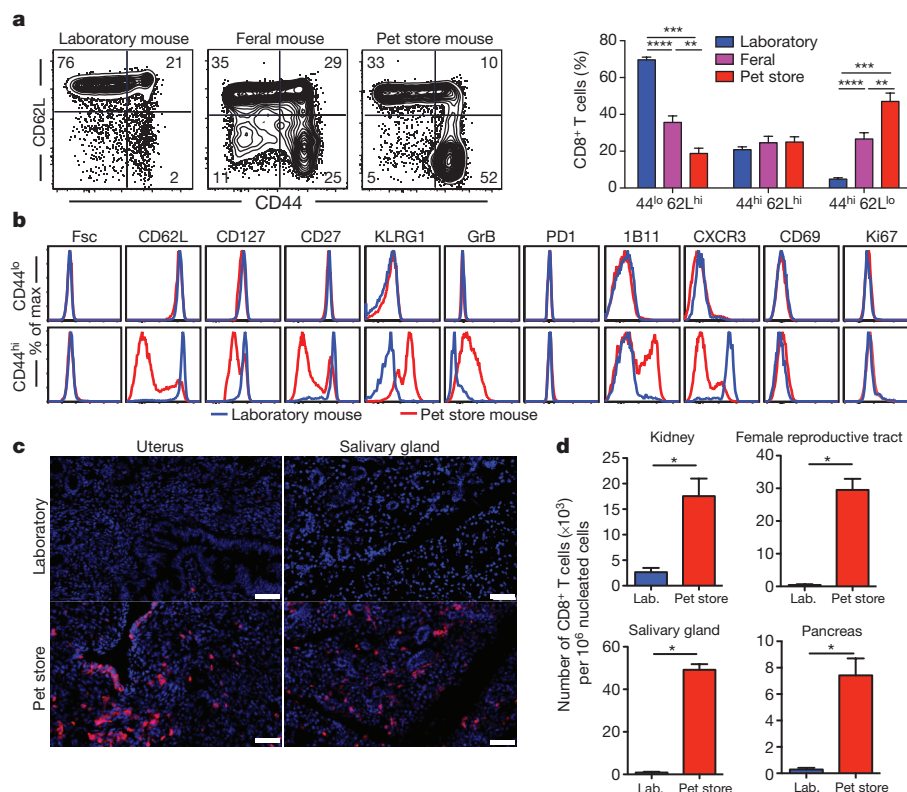


Figure 2 | CD8⁺ T-cell subsets vary among feral, pet store, and laboratory mice. **a**, CD8⁺ T-cell subsets in PBMCs from laboratory mice ($n=9$), feral mice that were trapped in the wild ($n=10$), and mice obtained from a pet store ($n=6$); results from two independent experiments. **b**, Phenotypes of CD44^{lo}/CD62L^{hi} (naive) and CD44^{hi} (antigen-experienced) CD8⁺ PBMCs from laboratory and pet store mice compared by fluorescence flow cytometry. **c**, **d**, CD8⁺ T-cell density

in non-lymphoid tissues from laboratory (Lab.) and pet store mice compared by quantitative immunofluorescence microscopy (QIM). Immunofluorescence staining of frozen sections of indicated tissues ($n=8$ animals per group; scale bars, 50 μ m; red, CD8 β ; blue, DAPI (nuclei)). Significance was determined using unpaired two-sided Mann–Whitney U -test. * $P<0.05$, ** $P<0.01$, *** $P<0.001$, **** $P<0.0001$; error bars indicate mean \pm s.e.m.

in fact, appeared much more similar to those of neonatal humans (Fig. 1c, d and Extended Data Fig. 1).

Mice are among the closest relatives of primates, but diverged from human ancestors 65–75 million years ago¹⁵. We wished to test whether the differences we observed between mice and adult humans were intrinsic to all mice (representing divergent immune system evolution) or unique to laboratory mice. To this end, we trapped free-living barn populations of feral mice and also procured mice from commercial pet stores. Compared to inbred SPF laboratory mice, feral and pet store mice had more antigen-experienced CD8⁺ T cells, particularly those that displayed a differentiated effector memory phenotype (Fig. 2a). More extensive phenotypic characterization of differentiation markers previously defined among CD8⁺ T cells in laboratory mice revealed that naive (CD44^{lo}) cells were phenotypically identical between laboratory and pet store mice (Fig. 2b). These data validated the consistency of markers used to discriminate between naive and antigen-experienced CD8⁺ T cell populations in the different mouse cohorts. However, examination of antigen-experienced (CD44^{hi}) CD8⁺ T cells indicated that pet store mice had more granzyme B⁺ and CD27^{lo} cells than did laboratory mice, and also showed increased signatures of terminal effector differentiation such as CD62L^{lo}, CXCR3^{lo}, CD127^{lo}, and 1B1^{hi} phenotypes. However, measures of cell size (Fsc) and proliferation (Ki67), and other markers of recent antigen exposure (CD69 and PD-1), were equivalent between both cohorts of mice, indicating that differences in CD8⁺ T-cell differentiation state were not due to ongoing acute infections in pet store mice. Moreover, the relative density of CD8⁺ T cells was up to 50-fold greater in nonlymphoid tissue from pet store mice than in tissue from laboratory mice (Fig. 2c, d).

Pet store mice recapitulated aspects of human CD8⁺ T-cell differentiation and distribution that were absent in laboratory mice. These

differences between mice could be solely due to genetics, or dependent on environment. Inbred mice are genetically homogeneous, which permits refined approaches and reductionist comparisons that are not possible in outbred populations, but they might have mutations that impair immune system development or function. Laboratory mice live in extremely hygienic filtered microisolator housing¹⁶. Conditions are so clean that mice with fatal immunodeficiency diseases often thrive in these environments owing to the absence of pathogen exposure. Thus, modern husbandry has evolved to the point in which laboratory mice, the standard model for biomedical research, acquire far less infectious experience than free-living (that is, ‘dirty’) mice or humans^{17–20}. To test whether environmental conditions might affect the differentiation state and distribution of CD8⁺ T cells, we co-housed inbred laboratory mice with pet store mice, which were not raised in ultra-hygienic barrier facilities. Adult pet store mice were introduced into cages containing adult laboratory (C57BL/6 strain) inbred mice. Within four weeks of co-housing, CD44^{hi} cells in the laboratory mice increased from 15% to approximately 70% of CD8⁺ peripheral blood mononuclear cells (PBMCs), and plateaued at around 50% after eight weeks for the duration of the study (Fig. 3a, b). During the first eight weeks of co-housing, 22% of laboratory mice died, but no further mortality was observed after this point (Fig. 3c). Serological tests for common mouse pathogens revealed exposure to viral, bacterial, and helminth pathogens (but not murine cytomegalovirus; see Extended Data Table 1). Co-housing resulted in a constitutive increase in highly differentiated effector memory cells in laboratory mice that matched the pattern seen in outbred mice and humans, including the accrual of granzyme B⁺ and CD27^{lo} cells (Fig. 3d). Moreover, nonlymphoid tissues in laboratory mice became populated by CD8⁺ T cells expressing a T_{RM}-cell phenotype (Fig. 3e and Extended Data Fig. 2). Expanding the cellular analysis

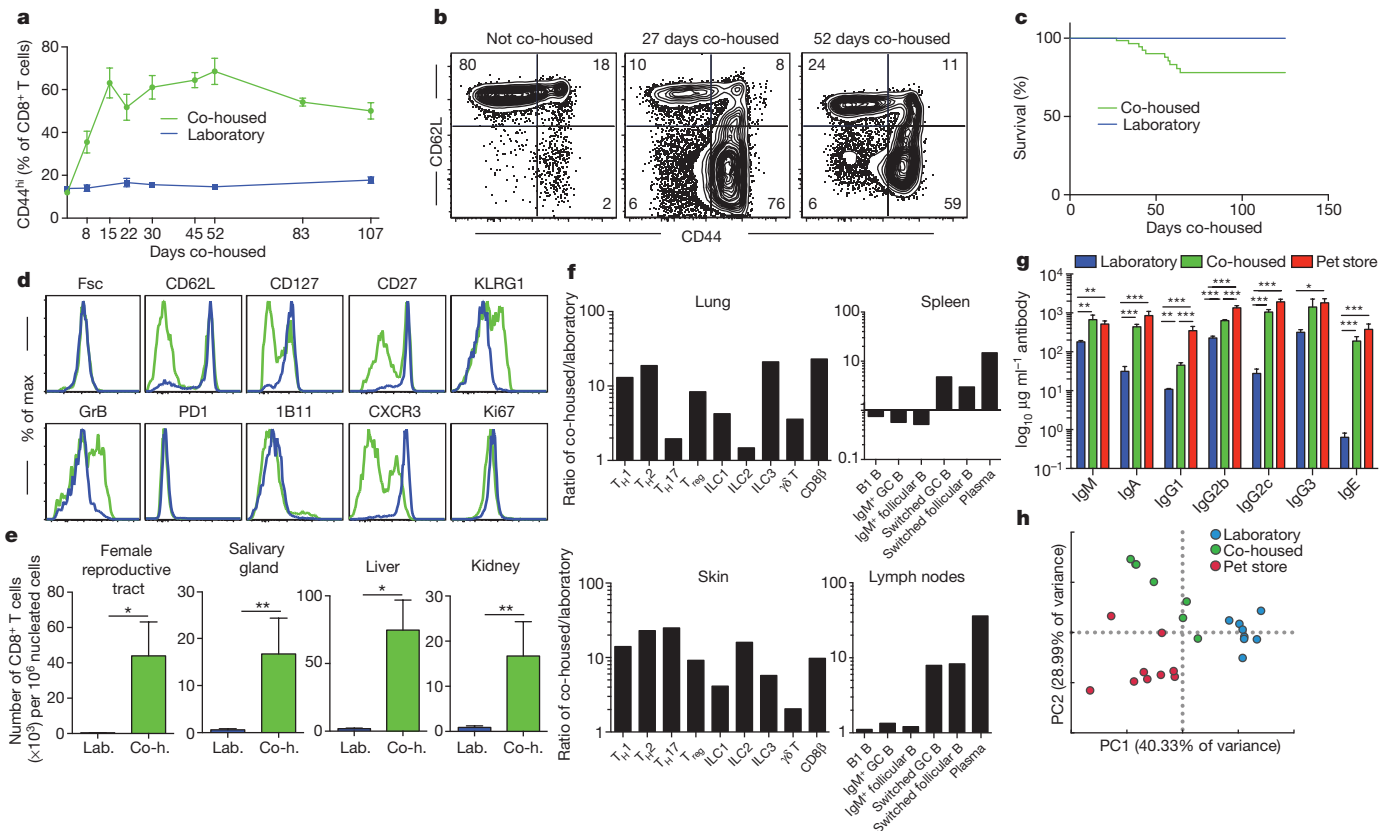


Figure 3 | Co-housing with pet store mice changes the immune system of laboratory mice. **a**, Proportion of CD44^{hi} (antigen-experienced) CD8⁺ PBMCs in laboratory mice housed in SPF conditions (blue, $n=10$ for each time point) or co-housed with pet store mice (green, $n=9$ for each time point) from two independent experiments. **b**, CD8⁺ T-cell phenotypes in blood from SPF or co-housed laboratory mice. Representative flow plots ($n=15$) are shown. **c**, Survival of SPF and co-housed laboratory mice ($n=26$ SPF and 65 co-housed mice). **d**, Representative phenotypes of CD44^{hi} CD8⁺ PBMCs isolated from laboratory mice co-housed with pet store mice for 100 days ($n=15$) and non-co-housed age-matched controls ($n=10$) were compared by fluorescence flow cytometry. **e**, Enumeration of CD8⁺ T cells in non-lymphoid tissues between laboratory and co-housed mice by QIM ($n=8$ animals per group).

beyond CD8⁺ T cells revealed extensive and profound changes to many innate and adaptive immune cell lineages in diverse tissues from co-housed laboratory mice and increased levels of serum antibodies (Fig. 3f, g and Extended Data Fig. 2). Furthermore, principal component analysis of PBMC gene expression data in the space of all detected genes (~18,000) revealed a spatial shift of co-housed samples away from laboratory samples and towards pet store samples along the first principal component (PC1; Fig. 3h). Together, these data demonstrate that co-housing profoundly altered the status of the immune system.

Our data suggested that the immune systems of laboratory mice have features in common with those of neonatal humans, and that altering the environment reproduced phenotypic immune signatures of adult humans. To test this hypothesis more broadly, we queried expression-profiling data from maternal and neonatal cord PBMCs from an unaffiliated study²¹ with pet store versus laboratory and co-housed versus laboratory mouse signatures using gene set enrichment analysis (GSEA; Fig. 4a, b). The top 400 genes that were upregulated in pet store mice showed highly significant enrichment in adult human expression data, whereas the top 400 downregulated genes showed enrichment in neonatal humans. Laboratory mice acquired this gene expression program after co-housing.

To more deeply investigate similarities in transcriptional patterns among pet store mice, co-housed mice, and adult humans compared

with laboratory mice and neonatal humans, we applied GSEA with the ImmuneSigDB database of immunological signatures²². We then used leading-edge metagene analysis of GSEA results to identify modules of coregulated genes that were upregulated in human adult versus cord PBMCs, and those found in pairwise comparisons of laboratory, pet store, and co-housed mice. Overlap between the resulting metagenes was used to identify global similarities between each data set. We observed highly significant overlaps between metagenes that were upregulated in adult PBMCs compared with cord PBMCs and metagenes that were upregulated in PBMCs from pet store or co-housed mice compared with laboratory mice. These included numerous pathways related to innate and adaptive immune functions (Fig. 4c, Extended Data Fig. 3 and Supplementary Tables 1 and 2). Conversely, metagenes that were upregulated in human cord blood cells overlapped with those upregulated in laboratory mice. Thus, these functional modules represent a major axis of similarity in immune status between pet store mice and adult humans relative to laboratory mice and neonatal humans, and can be conferred on laboratory mice through co-housing with pet store mice.

We next tested whether mouse husbandry affected immune responses. We challenged mice with the intracellular pathogen *Listeria monocytogenes*, a bacterial infection that is often used to gauge immune function in laboratory mice. Compared to laboratory mice, both pet

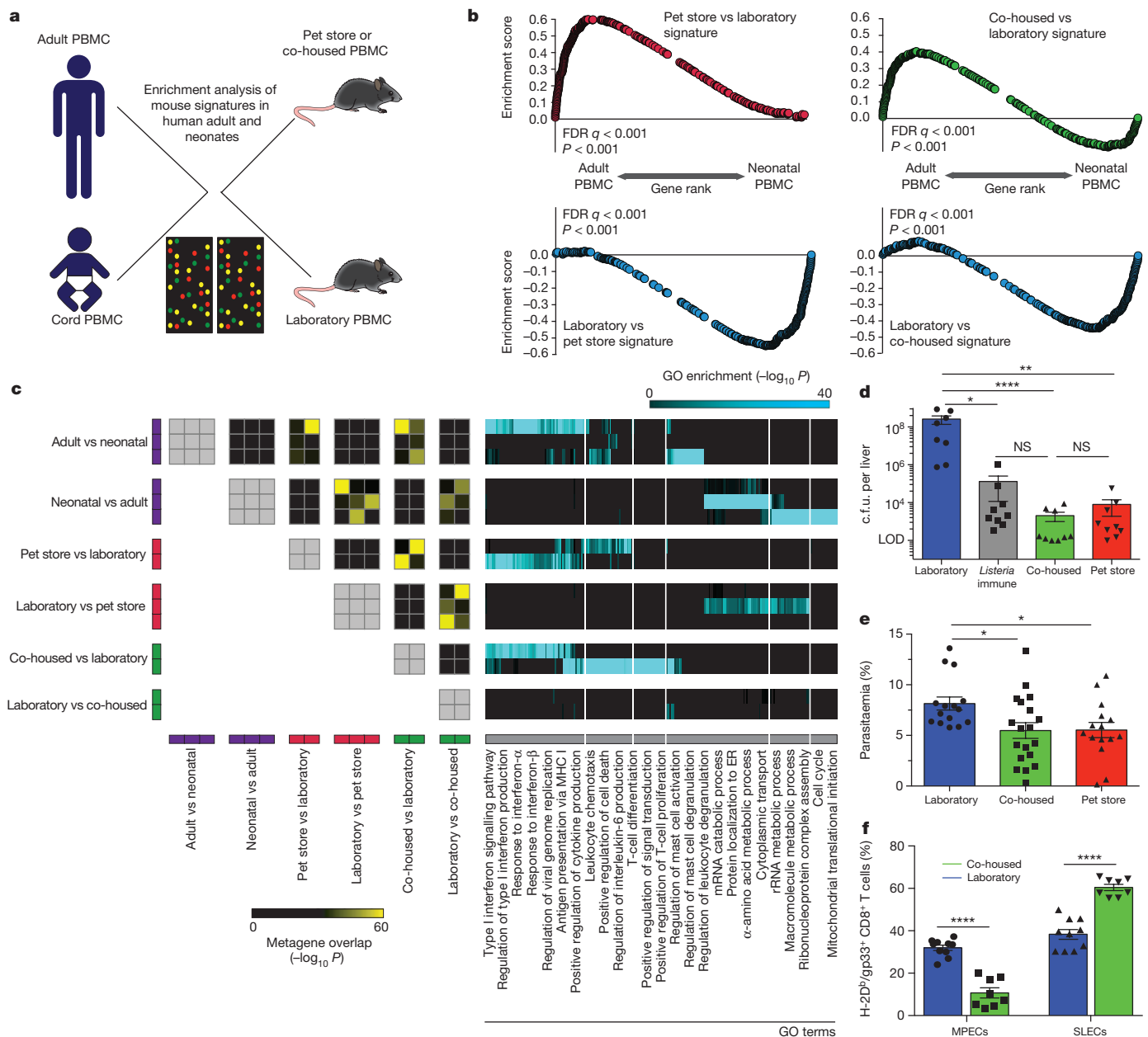


Figure 4 | Microbial experience matures mouse immune transcriptome from neonatal to adult human-like and affects immune system function. **a**, **b**, Enrichment of gene signatures among the indicated mouse group comparisons relative to human adult versus neonatal comparison. **a**, Experimental design. **b**, GSEA plots. Signatures consist of top 400 genes that were significantly differentially expressed in laboratory ($n = 8$), co-housed ($n = 7$), and pet store ($n = 8$) mice. **c**, Pairwise overlaps of metagenes identified through leading-edge metagene analysis and corresponding Gene Ontology (GO) terms. **d**, Bacterial load in the liver 3 days after challenge with 8.5×10^4 colony-forming units (c.f.u.) of *L. monocytogenes* ($n = 9$ for all except laboratory mice, $n = 8$). **e**, Parasitic

load in peripheral blood 5 days after *P. berghei* ANKA parasitized red blood cell challenge in laboratory ($n = 15$), co-housed ($n = 19$), and pet store ($n = 15$) mice. **f**, Twenty-eight days after LCMV infection (Armstrong strain), the proportion of H-2D^b/gp33-specific CD8⁺ T-cell MPECs (KLRG1⁺, CD127⁺) and SLECs (KLRG1⁺, CD127⁻) in PBMC from co-housed ($n = 8$) and laboratory ($n = 9$) mice. Cumulative data from two independent experiments. Data points in **d**–**f** represent individual mice. Significance was determined using Kruskal–Wallis (ANOVA) test (**d**), one-way ANOVA test (**e**) and unpaired two-sided *t*-test (**f**). * $P < 0.05$, ** $P < 0.01$, **** $P < 0.0001$; error bars indicate mean \pm s.e.m. NS, not significant. FDR, false discovery rate.

store and co-housed mice exhibited a $>10,000$ -fold reduction in bacterial burden three days after the challenge, and this matched bacterial control in laboratory mice that had been previously vaccinated against *L. monocytogenes* (Fig. 4d and Extended Data Fig. 4a). Hence, C57BL/6 mice that had experienced physiological exposure to environmental microbes exhibited considerably more innate resistance to *L. monocytogenes* infection than indicated from studies using SPF laboratory mice. The effect of mouse husbandry on infection control extended to a cerebral malaria model (*Plasmodium berghei* ANKA; Fig. 4e and Extended Data Fig. 4b). Lymphocytic choriomeningitis virus infection (LCMV)

is often used to investigate critical aspects of adaptive T-cell differentiation, including the regulation of memory precursor versus short-lived effector cell (MPEC versus SLEC) development^{23,24}. We observed that the proportions of MPECs and SLECs in LCMV-infected mice were significantly altered by mouse husbandry (Fig. 4f and Extended Data Fig. 4c).

Experiments in mice have informed much of our understanding of immune regulation, and have directly contributed to the development of life-saving clinical therapies. However, our study reveals an unanticipated impact of SPF husbandry on the immune system. Our results

do not support an end to SPF studies. However, it is ironic that such an immunologically inexperienced organism has become *de rigueur* for studies of the immune system, as our data show that this compromises development of a human adult-like immune system. To maximize opportunities to translate novel treatments from preclinical studies to clinical therapies, it may be opportune to add 'dirty' mice to our repertoire of investigative tools. Much as the analysis of truly sterile 'germ-free' mice has revealed how influential commensal flora are on 'normal' physiology and immune system function, our study suggests that the immune system in mice may not be fully 'normalized' without more complete microbial exposure¹⁰. Indeed, just as many autoimmune diseases do not manifest in genetically predisposed mice in the absence of commensal flora, certain infectious experiences have been shown to induce heterologous and innate immune memory, trigger autoimmune disease, and affect transplantation tolerance^{25–30}. Forward genetic screens to reveal the function of immunological genes are ongoing in mice, and it might be beneficial to conduct these screens in a dirty mouse model.

More generally, dirty mice might be valuable for investigating aspects of the hygiene hypothesis, immune function and treatments for disease in the settings of transplantation, allergy, autoimmunity, and vaccination, and perhaps in disparate diseases that might involve the immune or inflammatory systems (such as cardiovascular disease and cancer)⁴. Such mice could supplement current models to either increase translational potential to human disease or to better inform the efficacy of preclinical prophylactic and therapeutic modalities, without sacrificing powerful experimental tools and approaches that cannot be used in human studies.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 January; accepted 11 March 2016.

Published online 20 April 2016.

- Mestas, J. & Hughes, C. C. W. Of mice and not men: differences between mouse and human immunology. *J. Immunol.* **172**, 2731–2738 (2004).
- Seok, J. *et al.* Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proc. Natl Acad. Sci. USA* **110**, 3507–3512 (2013).
- Shay, T. *et al.* Conservation and divergence in the transcriptional programs of the human and mouse immune systems. *Proc. Natl Acad. Sci. USA* **110**, 2946–2951 (2013).
- Mak, I. W., Evaniew, N. & Ghert, M. Lost in translation: animal models and clinical trials in cancer treatment. *Am. J. Transl. Res.* **6**, 114–118 (2014).
- Rivera, J. & Tessarollo, L. Genetic background and the dilemma of translating mouse studies to humans. *Immunity* **28**, 1–4 (2008).
- Payne, K. J. & Crooks, G. M. Immune-cell lineage commitment: translation from mice to humans. *Immunity* **26**, 674–677 (2007).
- von Herrath, M. G. & Nepom, G. T. Lost in translation: barriers to implementing clinical immunotherapeutics for autoimmunity. *J. Exp. Med.* **202**, 1159–1162 (2005).
- Takao, K. & Miyakawa, T. Genomic responses in mouse models greatly mimic human inflammatory diseases. *Proc. Natl Acad. Sci. USA* **112**, 1167–1172 (2015).
- Schenkel, J. M. & Masopust, D. Tissue-resident memory T cells. *Immunity* **41**, 886–897 (2014).
- Thome, J. J. C. *et al.* Early-life compartmentalization of human T cell differentiation and regulatory function in mucosal and lymphoid tissues. *Nature Med.* (2016).
- Thome, J. J. C. *et al.* Spatial map of human T cell compartmentalization and maintenance over decades of life. *Cell* **159**, 814–828 (2014).
- Machado, C. S., Rodrigues, M. A. & Maffei, H. V. Gut intraepithelial lymphocyte counts in neonates, infants and children. *Acta Paediatr.* **83**, 1264–1267 (1994).
- Sallusto, F., Geginat, J. & Lanzavecchia, A. Central memory and effector memory T cell subsets: function, generation, and maintenance. *Annu. Rev. Immunol.* **22**, 745–763 (2004).
- Olson, J. A., McDonald-Hyman, C., Jameson, S. C. & Hamilton, S. E. Effector-like CD8⁺ T cells in the memory population mediate potent protective immunity. *Immunity* **38**, 1250–1260 (2013).
- Mouse Genome Sequencing Consortium Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
- Pritchett-Corning, K. R., Cosentino, J. & Clifford, C. B. Contemporary prevalence of infectious agents in laboratory mice and rats. *Lab. Anim.* **43**, 165–173 (2009).
- Pedersen, A. B. & Babayan, S. A. Wild immunology. *Mol. Ecol.* **20**, 872–880 (2011).
- Maizels, R. M. & Nussey, D. H. Into the wild: digging at immunology's evolutionary roots. *Nature Immunol.* **14**, 879–883 (2013).
- Cadwell, K. The virome in host health and disease. *Immunity* **42**, 805–813 (2015).
- Virgin, H. W., Wherry, E. J. & Ahmed, R. Redefining chronic viral infection. *Cell* **138**, 30–50 (2009).
- Votavova, H. *et al.* Transcriptome alterations in maternal and fetal cells induced by tobacco smoke. *Placenta* **32**, 763–770 (2011).
- Godec, J. *et al.* Compendium of immune signatures identifies conserved and species-specific biology in response to inflammation. *Immunity* **44**, 194–206 (2016).
- Kaech, S. M. *et al.* Selective expression of the interleukin 7 receptor identifies effector CD8 T cells that give rise to long-lived memory cells. *Nature Immunol.* **4**, 1191–1198 (2003).
- Joshi, N. S. *et al.* Inflammation directs memory precursor and short-lived effector CD8⁺ T cell fates via the graded expression of T-bet transcription factor. *Immunity* **27**, 281–295 (2007).
- Jordan, M. B., Hildeman, D., Kappler, J. & Marrack, P. An animal model of hemophagocytic lymphohistiocytosis (HLH): CD8⁺ T cells and interferon gamma are essential for the disorder. *Blood* **104**, 735–743 (2004).
- Selin, L. K. *et al.* Memory of mice and men: CD8⁺ T-cell cross-reactivity and heterologous immunity. *Immunol. Rev.* **211**, 164–181 (2006).
- Sun, J. C., Ugolini, S. & Vivier, E. Immunological memory within the innate immune system. *EMBO J.* **33**, 1295–1303 (2014).
- Adams, A. B., Pearson, T. C. & Larsen, C. P. Heterologous immunity: an overlooked barrier to tolerance. *Immunol. Rev.* **196**, 147–160 (2003).
- Taurog, J. D. *et al.* The germfree state prevents development of gut and joint inflammatory disease in HLA-B27 transgenic rats. *J. Exp. Med.* **180**, 2359–2364 (1994).
- Pozzilli, P., Signore, A., Williams, A. J. K. & Beales, P. E. NOD mouse colonies around the world: recent facts and figures. *Immunol. Today* **14**, 193–196 (1993).

Supplementary Information is available in the online version of the paper.

Acknowledgements This study was supported by National Institutes of Health grants 1R01AI111671, R01AI084913 (to D.M.), R01AI116678, R01AI075168 (to S.C.J.) and a BSL-3 suite rental waiver grant from the University of Minnesota. We thank R. Ahmed for providing reagents for pilot studies, P. Southern and D. McKenna for tissue samples or cord blood, and all members of the BSL-3 mouse team (University of Minnesota).

Author Contributions L.K.B., S.E.H., J.M.S., O.A.O., K.A.C., E.A.T., K.A.F., P.C.R., V.V., and D.M. performed the experiments and analysed the data. K.B. and W.N.H. analysed the transcriptome data. M.K.J., A.F.-M., and R.P.S. provided input on research design. L.K.B., S.E.H., W.N.H., S.C.J., and D.M. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.M. (masopust@umn.edu) or S.C.J. (james024@umn.edu).

METHODS

Mice, co-housing and infections. Pet store mice were purchased from various Twin Cities area pet stores. Feral mice were trapped on a horse farm or rural outdoor petting zoo in Minnesota or Georgia, USA. Male or female pet store mice were introduced into the cages of 6–8-week-old C57BL/6 mice of the same sex purchased from the National Cancer Institute. Co-housing occurred within a BSL-3 facility. Age-matched C57BL/6 laboratory mice maintained in SPF facilities served as controls. The number of animals needed to reach statistical significance was determined on the basis of previous experience. All animals that survived the experimental treatment were included in the final analysis. No method of randomization was used to allocate animals to experimental groups. Investigators were not blinded to the group allocation during experiments. *L. monocytogenes* was grown in tryptic soy broth containing streptomycin to log phase growth. The indicated groups of mice were infected intravenously (i.v.) with 8.5×10^4 c.f.u. of wild-type *L. monocytogenes* (provided by J. Harty). Bacterial load in the spleen and liver was determined 3 days post-challenge as previously described^{31,32}. *L. monocytogenes* immune mice were generated by primary infection with recombinant *L. monocytogenes* expressing OVA (LM-OVA) (provided by H. Shen)³³ 5 months before secondary challenge with wild-type *L. monocytogenes*. *P. berghei* ANKA (provided by S. K. Pierce) was propagated by passage in mice and blood collection. One-million parasitized RBCs were injected intraperitoneally (i.p.) into the indicated mice. Parasitaemia was measured by flow cytometry of peripheral blood³⁴. All mice were used in accordance with the guidelines of the Institutional Animal Care and Use Committees at the University of Minnesota.

Human tissue. Adult PBMC samples were collected from healthy volunteers at the University of Minnesota³⁵. Fresh cord blood samples were acquired from the Clinical Cell Therapy Laboratory at the University of Minnesota Medical Center. PBMC isolation has been described in detail elsewhere³⁶. After isolation, cells were frozen in 10^7 cells-per-ml aliquots in a cryopreservative solution (Sigma-Aldrich) for future phenotyping. Cervical tissue from premenopausal women was obtained from the Tissue Procurement Facility (BioNet, University of Minnesota). Cervical samples were frozen roughly 1–2 h after surgical resection. Informed consent was obtained from all subjects. The University of Minnesota Institutional Review Board approved all protocols used.

Intravascular staining, leukocyte isolation and phenotyping. An intravascular staining method was used to discriminate between cells present in the vasculature and cells in the tissue parenchyma³⁷. Briefly, animals were injected i.v. with biotin/fluorochrome-conjugated anti-CD45 through the tail vein. Three minutes after injection, animals were killed by cervical dislocation, and tissues were collected as described³⁸. Isolated mouse cells were surface-stained with antibodies against CD3 (145-2C11), CD45 (30F-11), CD11b (M1/70), CD11c (N418), NKp46 (29A1.4), Ly6G (1A8), MHC II (Ia-Ie) (M5/114.15.2), CD8 α (53-6.7), CD45.2 (104), CD4 (RM4-5), CD62L (MEL-14), CD44 (IM7), CD69 (H1.2F3), CD103 (M290), Ly6C (AL21), CD43 (1B11), CD43 (S7), CD27 (LG.3A10), PD-1 (RMP1-30 and J43), KLRG1 (2F1), CXCR3 (CXCR3-173), CD127 (SB/199), α 4 β 7 (DATK32), F4/80 (Cl⁻A3-1), CXCR5 (2G8), CD38 (90), IgM (RMM-1), IgD (11-26c.2a), GL7 (GL7), CD19 (6D5), and B220 (RA3-6B2). Isolated human cells were surface-stained with antibodies against CD8 α (3B5), CD45RA (HI100), CCR7 (G043H7), CD27 (O323), and CD3 (SK7). All of the above antibodies were purchased from BD Biosciences, Biolegend or Affymetrix eBiosciences. Cell viability was determined using Ghost Dye 780 (Tonbo Biosciences). Intracellular staining with phycoerythrin (PE)-conjugated granzyme B (Invitrogen), fluorescein isothiocyanate (FITC)-conjugated Ki67 (Invitrogen) and AF488-conjugated goat anti-mouse IgG (H+L) antibodies was performed using the Cytofix/Cytoperm kit (BD Pharmingen) following the manufacturer's instructions. Intracellular staining for transcription factors was performed using a transcription factor staining buffer set (Affymetrix eBiosciences) with antibodies against Foxp3 (FJK-16s), T-bet (4B10), Eomes (Dan11mag), Gata3 (LSO-823) and Ror γ t (Q31-378) following the manufacturer's guidelines. Single positive staining for T-bet, Gata3, and Ror γ t was used to identify T_H1, T_H2, and T_H17 lineages, respectively. FITC-conjugated mouse lineage cocktail (Tonbo Biosciences) was used in combination with other recommended lineage markers to identify various innate lymphoid cell subsets. The stained samples were acquired using LSRII or LSR Fortessa flow cytometers (BD) and analysed with FlowJo software (Tree Star, Inc.).

Infectious agent screening. Laboratory mice, co-housed laboratory mice (after at least 30 days of co-housing) and pet store mice were screened using EZ-spot and PCR Rodent Infectious Agent (PRIA) array methods (Charles River Laboratories). Dried whole blood, faeces, oral swabs and body swabs were collected as per the sample submission guidelines of Charles River Laboratories.

Tissue freezing, immunofluorescence and microscopy. Collected mouse tissues were fixed in 2% paraformaldehyde for 2 h before being treated with 30%

sucrose overnight for cryoprotection. The sucrose-treated tissue was embedded in tissue-freezing medium OCT and snap-frozen in an isopentane liquid bath. Human cervix specimens were embedded in tissue-freezing medium OCT and snap-frozen in an isopentane liquid bath. Frozen blocks were processed, stained, imaged, and enumerated by quantitative immunofluorescence microscopy (QIM) as described^{39,40}. Staining included the following antibodies: anti-mouse CD8 β (YTS156.7.7), anti-human CD8 β (SID8BEE, eBioscience), anti-mouse CD4 (RM4-5), anti-mouse CD11b (M1/70), and counterstaining with DAPI (mouse) or Cytox Green (human) to detect nuclei.

Serum antibody quantification. Mouse serum antibody titres were quantified using Ready-Set-Go! ELISA kits (Affymetrix eBioscience) following the manufacturer's instructions.

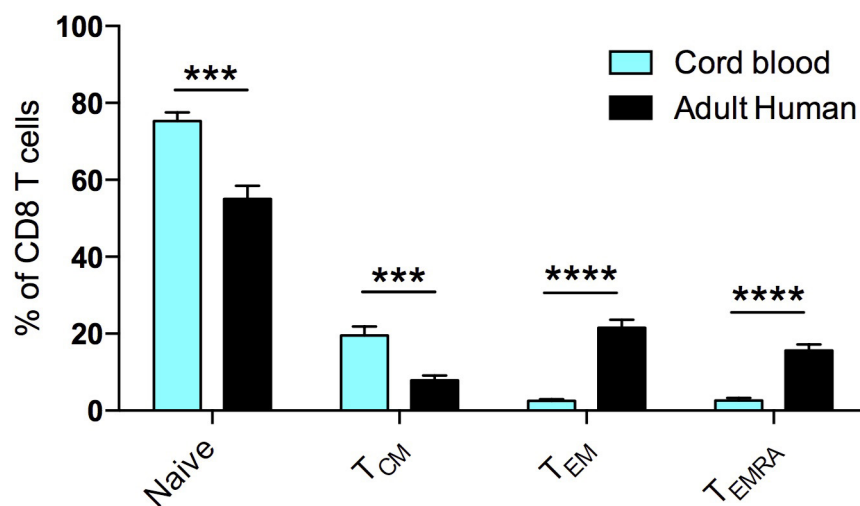
RNA isolation and microarray hybridization. For each sample, $1-3 \times 10^6$ PMBCs were used for RNA extraction. Cells were first homogenized using QIAshredder columns (Qiagen) and RNA was then extracted using an RNeasy kit (Qiagen) as per the manufacturer's instructions. Following quality control, total RNA samples were processed using the Illumina TotalPrep-96 RNA Amplification Kit for High-Throughput RNA Amplification for Array Analysis. Samples were loaded onto the MouseRef-8 v2.0 Expression BeadChip (Illumina) and hybridized beadchips were scanned using the Illumina iScan Beadarray Reader. Basic quality metrics were checked using Illumina Genomestudio.

Bioinformatics analysis. Before analysis, mouse microarray data were quantile normalized using preprocessCore (Bioconductor) and batch correction was performed using the ComBat algorithm. Principal components analysis was performed in R. Raw human adult and neonatal cord PBMC microarray data were obtained from a previous unaffiliated study profiling the peripheral blood of 72 smoking or non-smoking women and the cord blood of their neonates (Gene Expression Omnibus, accession code GSE27272)²¹. Human microarray data were quantile normalized as described above. To obtain lists of genes that were upregulated or downregulated among pet store, co-housed, and laboratory mice, differential expression analysis was performed using the linear modelling and empirical Bayesian method implemented in *limma* (Bioconductor). GSEA was performed as described previously⁴¹. LEM (Leading Edge Metagene) analysis was performed downstream of GSEA to yield groups of genes, termed metagenes, that are coordinately upregulated in a given phenotypic comparison and are common to multiple enriched gene sets. Briefly, for a given phenotypic comparison, GSEA was performed using ImmuneSigDB, a curated compendium of 4,872 gene sets describing a wide range of cell states and experimental perturbations from immunology literature²². The top 150 significantly enriched gene sets, as restricted by an FDR < 0.25 and ranked by $P < 0.05$, were subsetted for their leading edge genes. These genes were then clustered into metagenes using non-negative matrix factorization. The significance of overlap between pairs of metagenes was determined using a Fisher exact test ($P < 1 \times 10^{-5}$). Metagenes were functionally annotated based on the significance of overlap between member genes and GO terms⁴¹, as measured by hypergeometric test using the GOrilla enrichment analysis tool⁴².

Statistics. Data were subjected to the D'Agostino and Pearson omnibus normality test to determine whether they were sampled from a Gaussian distribution. If a Gaussian model of sampling was satisfied, parametric tests (unpaired two-tailed Student's *t*-test for two groups and one-way ANOVA with Bonferroni multiple comparison test for more than two groups) were used. If the samples deviated from a Gaussian distribution, non-parametric tests (Mann–Whitney U test for two groups, Kruskal–Wallis with Dunn's multiple comparison test for more than two groups) were used unless otherwise stated. Variances between groups were compared using an F test and found to be equal. All statistical analysis was done in GraphPad Prism (GraphPad Software Inc.). $P < 0.05$ was considered significant.

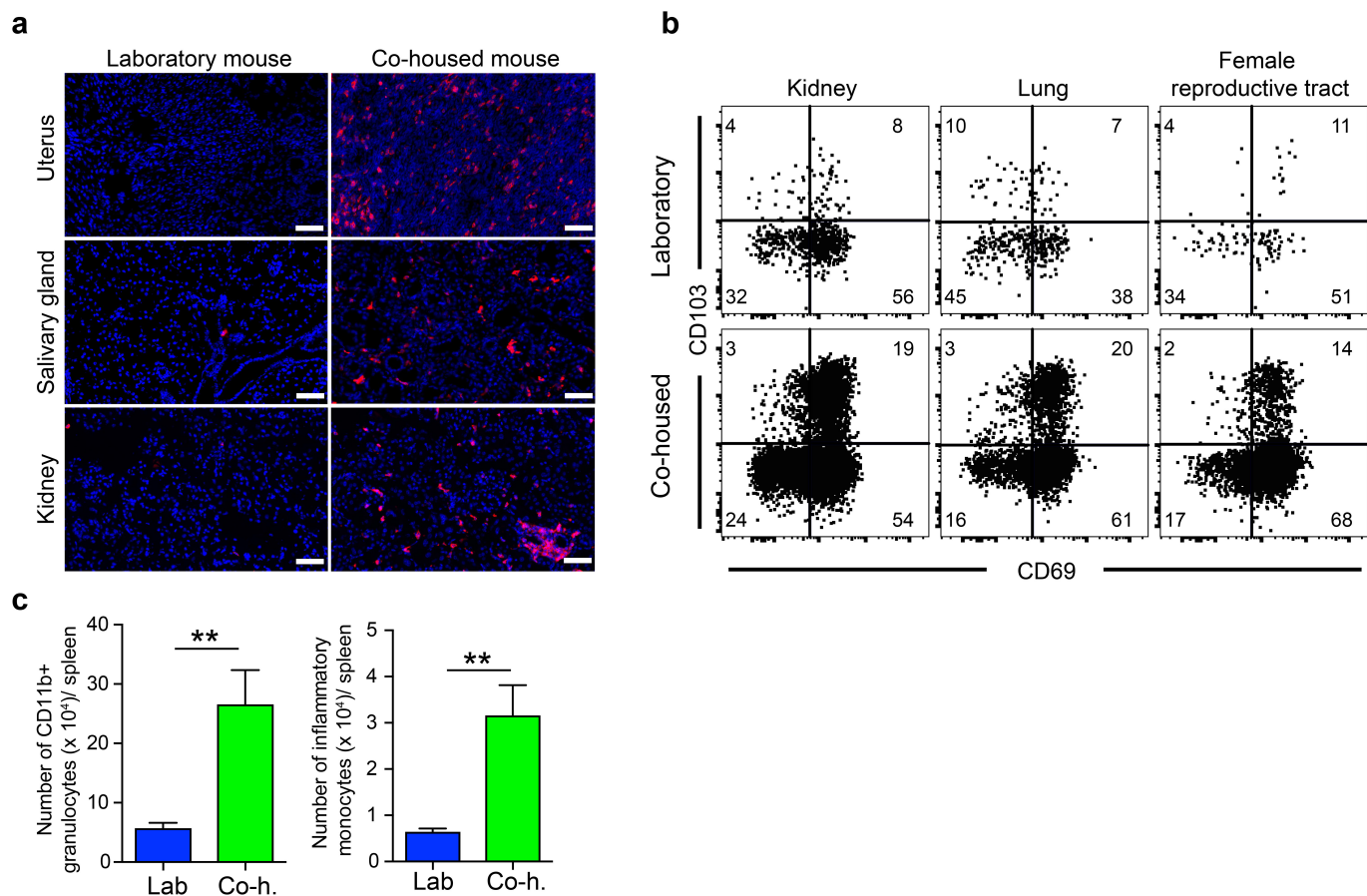
- Hamilton, S. E., Wolkers, M. C., Schoenberger, S. P. & Jameson, S. C. The generation of protective memory-like CD8⁺ T cells during homeostatic proliferation requires CD4⁺ T cells. *Nature Immunol.* **7**, 475–481 (2006).
- Hamilton, S. E., Schenkel, J. M., Akue, A. D. & Jameson, S. C. IL-2 complex treatment can protect naive mice from bacterial and viral infection. *J. Immunol.* **1950**, 6584–6590 (2010).
- Pope, C. et al. Organ-specific regulation of the CD8 T cell response to *Listeria monocytogenes* infection. *J. Immunol.* **1950**, 3402–3409 (2001).
- Gordon, E. B. et al. Inhibiting the Mammalian target of rapamycin blocks the development of experimental cerebral malaria. *MBio* **6**, e00725 (2015).
- Balfour, H. H. et al. Behavioral, virologic, and immunologic factors associated with acquisition and severity of primary Epstein–Barr virus infection in university students. *J. Infect. Dis.* **207**, 80–88 (2013).
- Odumade, O. A. et al. Primary Epstein–Barr virus infection does not erode preexisting CD8⁺ T cell memory in humans. *J. Exp. Med.* **209**, 471–478 (2012).

37. Anderson, K. G. *et al.* Intravascular staining for discrimination of vascular and tissue leukocytes. *Nature Protocols* **9**, 209–222 (2014).
38. Beura, L. K. *et al.* Lymphocytic choriomeningitis virus persistence promotes effector-like memory differentiation and enhances mucosal T cell distribution. *J. Leukoc. Biol.* **97**, 217–225 (2015).
39. Schenkel, J. M., Fraser, K. A., Vezys, V. & Masopust, D. Sensing and alarm function of resident memory CD8⁺ T cells. *Nature Immunol.* **14**, 509–513 (2013).
40. Steinert, E. M. *et al.* Quantifying memory CD8 T cells reveals regionalization of immunosurveillance. *Cell* **161**, 737–749 (2015).
41. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
42. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).



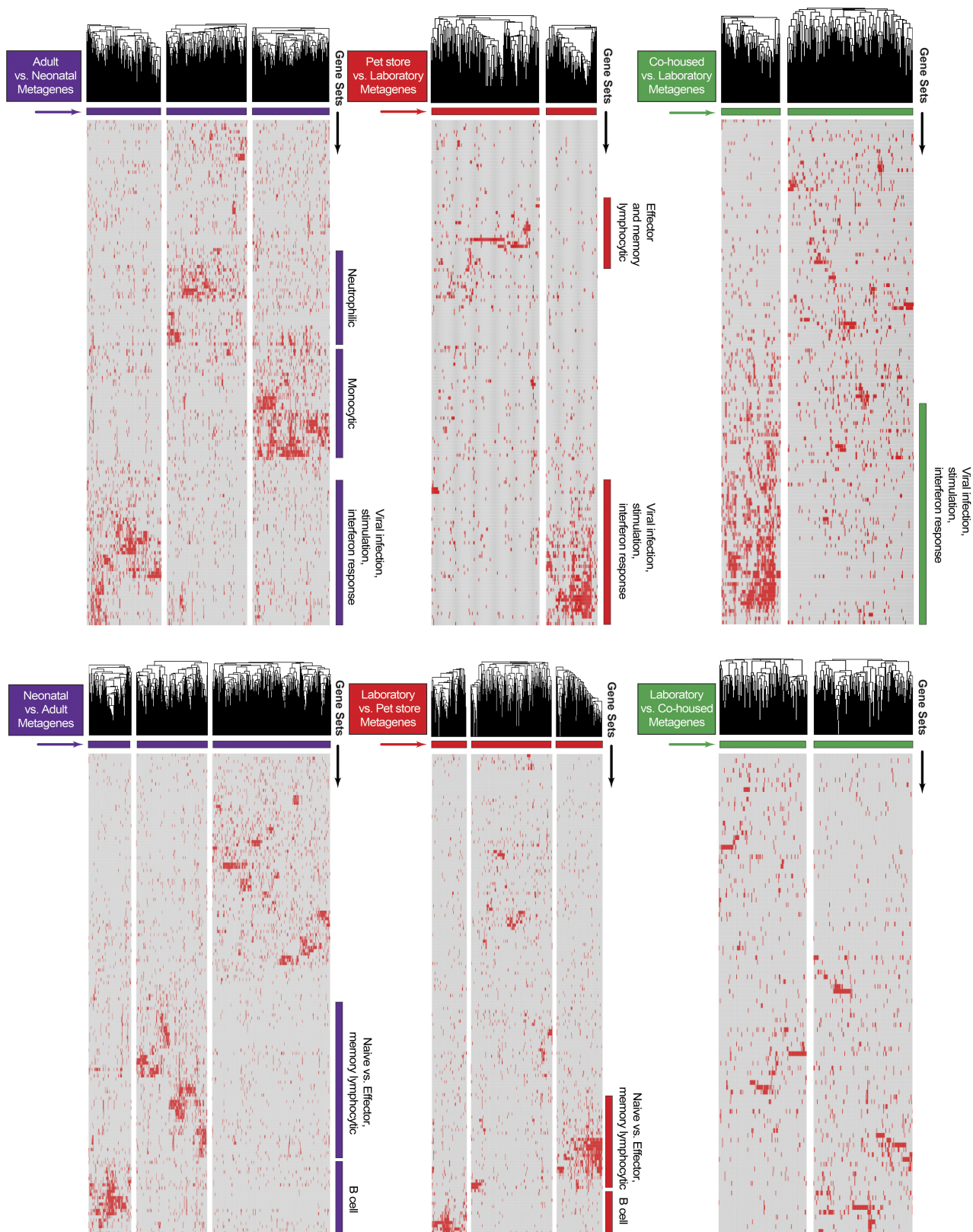
Extended Data Figure 1 | Frequency of CD8⁺ T-cell subsets in newborn versus adult humans. CD8⁺ T-cell subsets were defined in adult PBMCs ($n = 13$) and cord blood PBMCs ($n = 8$) by fluorescence flow cytometry based on the following markers: naive,

CD45RA^{hi}CCR7^{hi}; T_{CM}, CD45RA^{lo}CCR7^{hi}; T_{EM}, CD45RA^{lo}CCR7^{lo}; T_{EMRA}, CD45RA^{hi}CCR7^{lo}. Significance was determined using unpaired two-sided *t*-test. *** $P < 0.001$, **** $P < 0.0001$; error bars indicate mean \pm s.e.m.



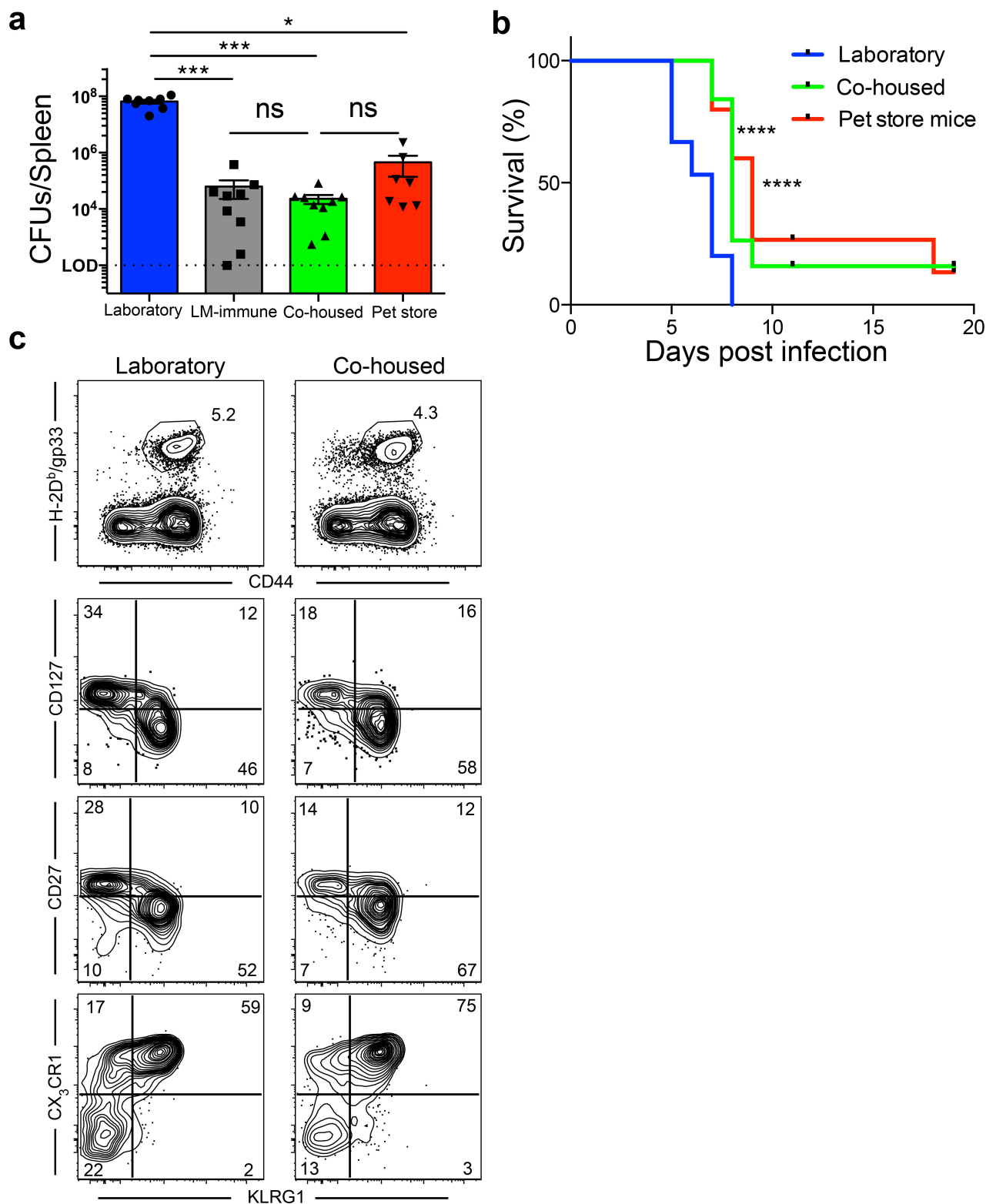
Extended Data Figure 2 | Co-housing laboratory mice with pet store mice induces accumulation of T_{RM} -phenotype $CD8^+$ T cells and other innate cells in tissues of laboratory mice. **a**, $CD8^+$ T-cell density within the indicated tissues of adult laboratory mice ($n=5$) and co-housed mice ($n=7$). Representative immunofluorescence staining, $CD8\beta$ (red), DAPI (nuclei, blue); scale bars, $50\ \mu m$. **b**, Phenotype of $CD8^+$ T cells was compared between laboratory mice ($n=9$) and age-matched laboratory

mice that were co-housed ($n=9$, representative flow cytometry plots shown). Samples gated on $CD44^{hi}$ cells isolated from the indicated tissue (vasculature populations were excluded, see Methods). **c**, Enumeration of $CD11b^+$ granulocytes and $Ly6C^{hi}$ inflammatory monocytes in spleens of laboratory ($n=6$) and co-housed ($n=6$) mice. Significance was determined using unpaired two-sided Mann–Whitney U -test. $^{**}P < 0.01$; error bars indicate mean \pm s.e.m.



Extended Data Figure 3 | LEM metagene analysis. For each comparison, standard GSEA was performed using the ImmSigDB database of gene-sets. Genes in the top 150 enriched sets ($FDR < 0.001$, ranked by P value) were filtered to only leading edge genes and subsequently clustered into groups (metagenes) using an NMF algorithm. Hierarchical clustering of genes within individual metagenes was performed to obtain the final heatmap. Metagenes with qualitatively discernible 'blocks' of gene-set membership

were annotated according to the identity of corresponding enriched gene-sets. Heatmaps for adult versus neonatal, pet store versus laboratory, co-housed versus laboratory, neonatal versus adult, laboratory versus pet store, and laboratory versus co-housed comparisons are shown. Individual genes within each metagene are listed in Supplementary Table 1. Pairwise overlaps between metagenes from different comparisons are visualized in Fig. 4c.



Extended Data Figure 4 | Environment altered antimicrobial resistance and CD8⁺ T-cell differentiation. Laboratory mice were co-housed with pet store mice as described in Figure 3. **a**, Bacterial load in the spleen 3 days after challenge with 8.5×10^4 c.f.u. of *L. monocytogenes* (LM) in laboratory ($n=8$), LM-immune ($n=9$), co-housed ($n=9$) and pet store mice ($n=9$) in two independent experiments. **b**, Survival of laboratory mice ($n=15$), co-housed mice ($n=19$) and pet store mice ($n=15$) after challenge with 10^6 *P. berghei* ANKA parasitized RBCs in two

independent experiments. **c**, Laboratory ($n=9$) and co-housed ($n=8$) mice were infected with LCMV. Four weeks later, LCMV-specific CD8⁺ T cells (identified with H-2D^b/gp33 MHC I tetramers) were evaluated for expression of the indicated markers. Top row, gated on live CD8⁺ T cells. Bottom three rows, gated on live CD8⁺ H-2D^b/gp33⁺ T cells. Significance was determined using Kruskal–Wallis (ANOVA) test (**a**) and log-rank (Mantel–Cox) test (**b**). * $P < 0.05$, *** $P < 0.001$, **** $P < 0.0001$; error bars indicate mean \pm s.e.m.

Extended Data Table 1 | Microbial exposure in laboratory, pet store and co-housed mice

	Pet store	Laboratory	Co-housed
Viruses			
Rotavirus (EDIM)	0	0	0
Mouse Hepatitis Virus	93.3	0	61.5
Murine norovirus	60	0	38.5
Mouse parvovirus NS1	53.3	0	0
Mouse parvovirus type 1	40	0	7.7
Mouse parvovirus type 2	46.7	0	0
Minute virus of mice	46.7	0	0
Theiler's murine encephalomyelitis virus	60	0	38.5
Sendai virus	66.7	0	23.1
Ectromelia virus	0	0	0
Lymphocytic Choriomeningitis virus	6.7	0	7.7
Mouse adenovirus 1 and 2	0	0	0
Mouse cytomegalo virus	0	0	0
Polyoma virus	6.7	0	0
Pneumonia virus of mouse	53.3	0	0
Reovirus	0	0	0
Bacteria			
Cillia-Associated Respiratory Bacillus	0	0	0
Mycoplasma pulmonis	73.3	0	30.8
Clostridium piliforme	26.7	0	0
Parasites/Protozoa/Fungi			
Encephalitozoon cuniculi	40	0	0
Pinworm	100	0	100
Mites	100	0	100

Frequency of each indicated microbial exposure within the population was evaluated by serological analysis and/or PCR in laboratory ($n = 4$), pet store ($n = 15$) and co-housed ($n = 13$) mice. Each co-housed sample was collected from a different cage. Pinworm- and mite-specific PCR was performed on pooled pet store and co-housed samples.

The CRISPR–associated DNA–cleaving enzyme Cpf1 also processes precursor CRISPR RNA

Ines Fonfara^{1,2,3*}, Hagen Richter^{2,3*}, Majda Bratovič^{2,3,4}, Anaïs Le Rhun^{1,2,3} & Emmanuelle Charpentier^{1,2,3,4}

CRISPR–Cas systems that provide defence against mobile genetic elements in bacteria and archaea have evolved a variety of mechanisms to target and cleave RNA or DNA¹. The well-studied types I, II and III utilize a set of distinct CRISPR-associated (Cas) proteins for production of mature CRISPR RNAs (crRNAs) and interference with invading nucleic acids. In types I and III, Cas6 or Cas5d cleaves precursor crRNA (pre-crRNA)^{2–5} and the mature crRNAs then guide a complex of Cas proteins (Cascade-Cas3, type I; Csm or Cmr, type III) to target and cleave invading DNA or RNA^{6–12}. In type II systems, RNase III cleaves pre-crRNA base-paired with *trans*-activating crRNA (tracrRNA) in the presence of Cas9 (refs 13, 14). The mature tracrRNA–crRNA duplex then guides Cas9 to cleave target DNA¹⁵. Here, we demonstrate a novel mechanism in CRISPR–Cas immunity. We show that type V–A Cpf1 from *Francisella novicida* is a dual-nuclease that is specific to crRNA biogenesis and target DNA interference. Cpf1 cleaves pre-crRNA upstream of a hairpin structure formed within the CRISPR repeats and thereby generates intermediate crRNAs that are processed further, leading to mature crRNAs. After recognition of a 5′-YTN-3′ protospacer adjacent motif on the non-target DNA strand and subsequent probing for an eight-nucleotide seed sequence, Cpf1, guided by the single mature repeat-spacer crRNA, introduces double-stranded breaks in the target DNA to generate a 5′ overhang¹⁶. The RNase and DNase activities of Cpf1 require sequence- and structure-specific binding to the hairpin of crRNA repeats. Cpf1 uses distinct active domains for both nuclease reactions and cleaves nucleic acids in the presence of magnesium or calcium. This study uncovers a new family of enzymes with specific dual endoribonuclease and endonuclease activities, and demonstrates that type V–A constitutes the most minimalistic of the CRISPR–Cas systems so far described.

Our previous analysis of the intracellular human pathogen *Francisella novicida* U112 by small RNA (sRNA) sequencing identified sRNAs expressed from two CRISPR–Cas loci^{13,16} (Extended Data Fig. 1a). As well as for the type II–B locus¹³, we detected sRNAs from a CRISPR–Cas locus that resembled the minimal architecture of type II systems but lacked a *cas9* gene. Upstream of the *cas1*, *cas2* and *cas4* genes¹⁷, FTN_1397 was identified as a *cas* gene encoding a protein distinct in sequence from known Cas proteins; this was later named *cpf1* (*cas* gene of *Pasteurella*, *Francisella*)¹⁷. This system was recently classified as a type V–A system belonging to the class 2 CRISPR–Cas systems^{18,19}. The CRISPR array contains a series of nine spacer sequences separated by 36-nucleotide (nt) repeat sequences. The mature RNAs are composed of a repeat sequence in 5′ and spacer sequence in 3′, similar to the repeat-spacer composition of types I and III systems but distinct from the spacer-repeat composition of type II systems^{2,14,20} (Extended Data Fig. 1b). As in type I, the repeat forms a hairpin structure at its 3′ end²⁰. Neither the presence of a Cas6 homologue nor the expression of a tracrRNA-like sRNA could be detected in the vicinity of the

F. novicida type V–A locus, indicating that Cpf1 uses a distinct mode of crRNA biogenesis compared to the mechanisms that have been described thus far^{2,4,14}.

We investigated whether Cpf1 acts as the single effector enzyme in pre-crRNA processing in type V–A systems. Recombinant *F. novicida* Cpf1 protein was overexpressed, purified and biochemically characterized. In contrast to the recently reported formation of Cpf1 dimers in solution¹⁶, our data reveal a molecular weight of 187 kDa (Extended Data Fig. 2), indicating that Cpf1 is a monomer. This result is corroborated by another study showing the crystal structure of Cpf1 from *Lachnospiraceae* bacterium (LbCpf1). No oligomerization of Cpf1 was observed in the crystals, analytical ultracentrifugation experiments or electron microscopy²¹. The monomeric nature is consistent with Cpf1 forming a complex with the guide crRNA to bind and cleave target DNA because if the active protein was a dimer¹⁶, it would probably require a tandem DNA target site, or alternatively, two different crRNAs targeting the top and bottom strand of the DNA.

In vitro cleavage assays show that Cpf1 processes a pre-crRNA consisting of a full-length repeat-spacer, yielding a 19-nt repeat fragment, and a 50-nt repeat-spacer crRNA intermediate (Fig. 1). Only RNAs with full-length repeat sequences were processed, indicating that the RNA cleavage activity is repeat-dependent (Extended Data Fig. 3a). The observed cleavage site is in good agreement with the data obtained by RNA-seq (Extended Data Fig. 1b) and a recent study¹⁶. The crRNAs produced *in vitro* represent intermediate forms that undergo further processing at the 5′ and 3′ ends by a nonspecific mechanism *in vivo*. Cpf1 cleaves pre-crRNA four nucleotides upstream of the stem-loop (Fig. 1). The cleavage site is reminiscent of many Cas6 enzymes and Cas5d, which recognize the hairpin of their respective repeats^{2,4,5,20}. Cpf1, however, does not cleave directly at the base of the stem-loop, suggesting that the structure is not the only requirement for processing of pre-crRNA. Northern blot analysis using an inducible *Escherichia coli* heterologous system also demonstrates processing of pre-crRNA upon Cpf1 expression (Extended Data Fig. 3b), resulting in the expected RNA fragments.

To investigate the importance of the repeat and its hairpin structure in successful Cpf1 processing, we designed RNAs with mutations that yield either an altered repeat sequence keeping the stem-loop structure or an unstructured repeat. In contrast to the wild-type RNA substrate containing an intact repeat, none of the mutated RNAs was cleaved by Cpf1 (Extended Data Fig. 4a, b). We further designed repeat variants with either single nucleotide mutations between the cleavage site and the stem-loop (a region referred to as repeat recognition sequence (RRS)) or different sizes of the loop and stem regions (Extended Data Fig. 4a). Single nucleotide mutations in the RRS yielded repeat variants that were not, or only poorly, cleaved by Cpf1 (Extended Data Fig. 4c), indicating that these residues between the stem and the cleavage site have a role in processing of the substrate. This can be explained by the distinct secondary structure of crRNA in complex with Cpf1, where

¹The Laboratory for Molecular Infection Medicine Sweden (MIMS), Umeå Centre for Microbial Research (UCMR), Department of Molecular Biology, Umeå University, Umeå 90187, Sweden.

²Helmholtz Centre for Infection Research, Department of Regulation in Infection Biology, Braunschweig 38124, Germany. ³Max Planck Institute for Infection Biology, Department of Regulation in Infection Biology, Berlin 10117, Germany. ⁴Hannover Medical School, Hannover 30625, Germany.

*These authors contributed equally to this work.

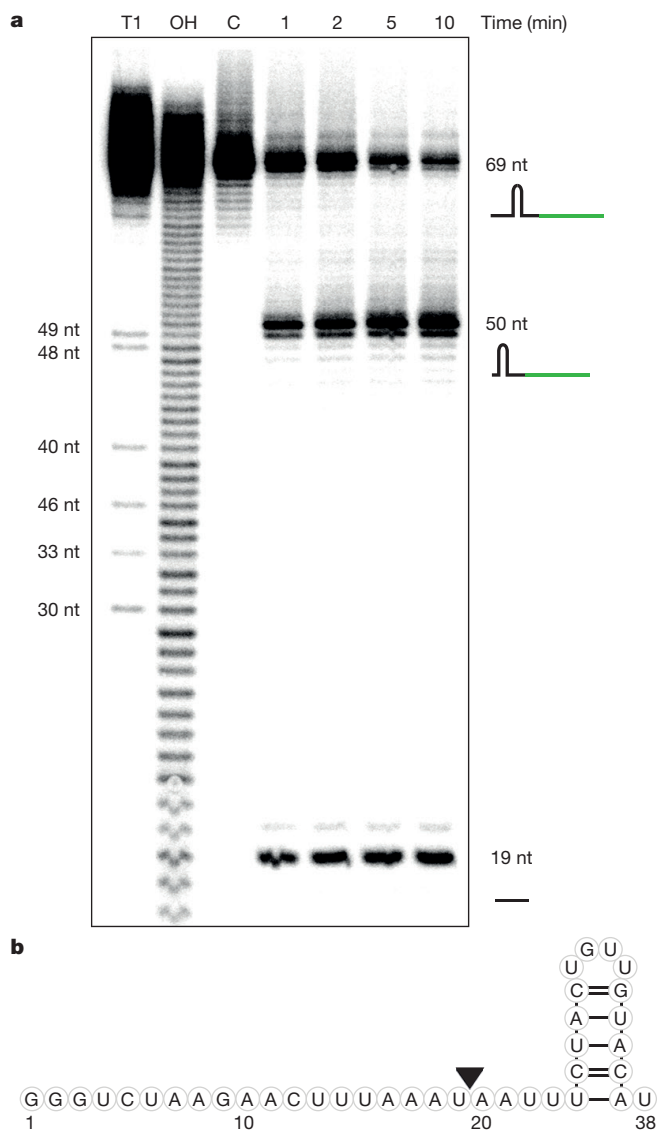


Figure 1 | Cpf1 processes pre-crRNA upstream of the repeat stem-loop structure. **a**, Denaturing polyacrylamide gel showing the processing of internally labelled 69-nt pre-crRNA (200 nM) by Cpf1 (1 μ M) in the presence of 10 mM $MgCl_2$ over 10 min. T1, RNase T1 ladder; OH, alkaline hydrolysis ladder; C, control reaction without Cpf1. Shown is a representative of three independent experiments. **b**, Schematic representation of pre-crRNA repeat structure. The Cpf1 cleavage site is indicated by a black triangle.

the RRS folds back to make contacts with the stem-loop²¹. Changes in the loop region of the repeat structure resulted in reduced cleavage activity for a shorter loop, whereas an increased loop length did not influence cleavage (Extended Data Fig. 4d). Extensive contacts of Cpf1 to the stem-loop of the crRNA²¹ explain why alterations of the stem structure yielded non-cleavable substrates. These results highlight the requirement of a stem-loop structure specific in length and sequence for recognition by Cpf1. Thus, the repeat cleavage reaction is highly sequence- and structure-dependent.

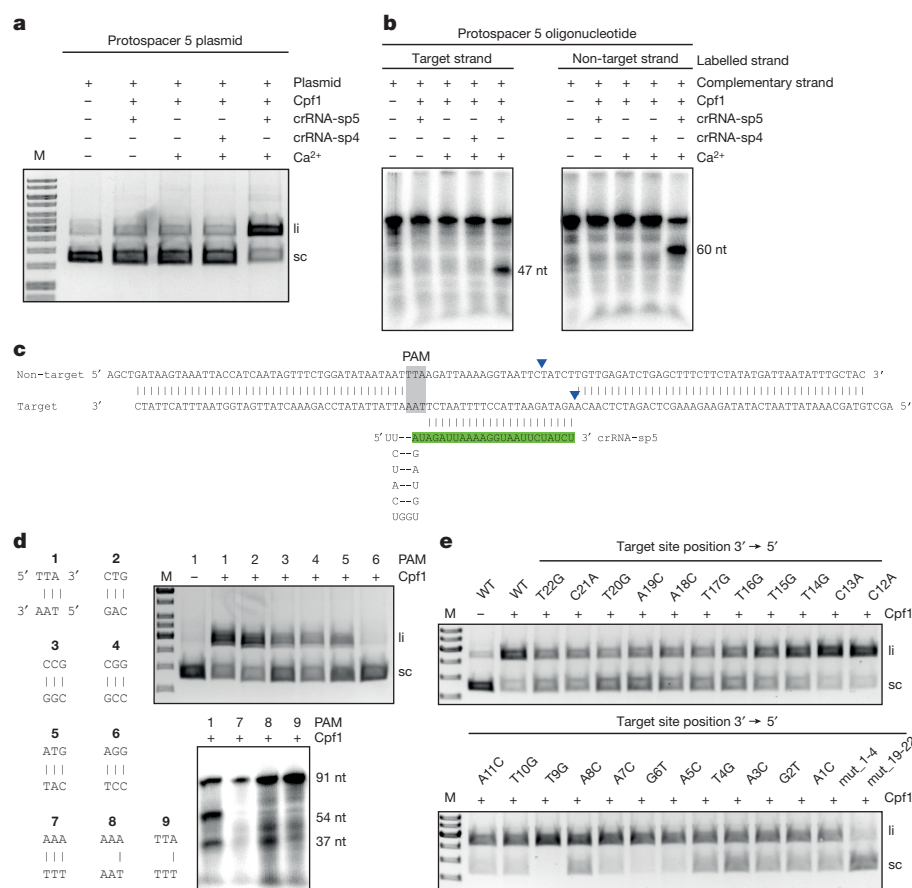
To determine the ion dependency of Cpf1 processing activity, we tested a variety of divalent metal ions in RNA cleavage assays. The activity of Cpf1 in pre-crRNA processing was highest when Mg^{2+} was added to the reaction (Extended Data Fig. 5a). Addition of Ca^{2+} , Mn^{2+} or Co^{2+} also mediated cleavage, although not to the level of specificity observed with Mg^{2+} . Equimolar addition of EDTA markedly reduced Cpf1 processing activity. The dependency on Mg^{2+} is in contrast to the ion-independent reaction of Cas6 (types I and III)^{2,20} or Cas5d

(type I-C)⁵. A Mg^{2+} ion is coordinated in the structure of the crRNA²¹. Whether this ion is required for catalysis or only for stabilization of the tertiary structure has not yet been determined. Thus, our study highlights a novel crRNA biogenesis mechanism in which Cpf1 is a metal-dependent endoribonuclease that cleaves pre-crRNA in a sequence- and structure-specific manner. Similarities in the pre-crRNA processing mechanisms of Cpf1 and Cas6 enzymes of type I and type III systems indicate potential evolution of these ancestral CRISPR–Cas systems through transposition events¹⁸. This hypothesis is supported by our finding that Cpf1 functions as the endoribonuclease of type V-A systems together with the repeat-spacer composition of mature crRNAs and the requirement for a hairpin structure in the repeat. Bioinformatic analyses indicate that type V systems may be ancestral versions of type II systems. Type V may be considered as a link between class 1 and class 2 systems, which is supported by the recent discovery of a subtype V-B that encodes tracrRNA^{18,19}.

It was previously shown that Cpf1 acts as the DNA endonuclease guided by crRNA to cleave double-stranded (ds)DNA site-specifically¹⁶. In accordance with that study, we show that only crRNA containing an intact stem-loop and a sequence complementary to the target DNA mediated Cpf1 DNA cleavage that resulted in a staggered cut producing a 5-nt 5' overhang (Fig. 2a, b; processed crRNAs (RNA1–3), full-length pre-crRNAs (RNA4–6), mutated crRNAs (RNA7 and 8), Extended Data Figs 6 and 7). Surprisingly, a crRNA with a spacer-repeat arrangement also mediated cleavage by Cpf1, albeit with less efficiency than the wild type. Although the RNA processing activity of Cpf1 is highly dependent on the repeat sequence (sequence mutant, Extended Data Fig. 4a, b), a similar RNA resulted in residual DNA cleavage activity (RNA7, Extended Fig. 6). This might be due to the 3' end nucleotide of the repeat, which was not mutated and was recently reported to be crucial for DNA targeting¹⁶ and for maintaining the specific tertiary structure of crRNA²¹.

Given that Cpf1 can process pre-crRNA, it is not surprising that RNAs with the full-length repeat-spacer (RNA4 and RNA6, Extended Data Fig. 6) mediate similar cleavage activities as the mature crRNA form. RNA containing the full-length repeat-spacer led to the most efficient DNA binding and nuclease activity of Cpf1 (compare RNA4 to RNA3 and RNA6, Extended Data Figs 8a and 6a, b). The processed form of crRNA (RNA3, Extended Data Fig. 6) was constructed on the basis of sRNA sequencing results (Extended Data Fig. 1) before the exact RNA processing of Cpf1 was known (Fig. 1), which resulted in a 3-nt shorter 5' end. Binding to and processing of pre-crRNA induces conformational changes in Cpf1, causing the enzyme to change into an active endonucleolytic state²¹. Similarly, an induced-fit mechanism is used by Cas9, which undergoes large conformational rearrangements upon binding to tracrRNA–crRNA²².

A seed sequence of 3–5 nt at the PAM-proximal side of the protospacer has been reported for Cpf1 (ref. 16). Using plasmids with single mismatches between spacer and protospacer along the target sequence, we observed that Cpf1 was sensitive to mismatches within the first eight PAM proximal nucleotides, and would not tolerate four consecutive mismatches. Furthermore, Cpf1 was sensitive to mismatches around the cleavage site (position 1–4 on the PAM-distal site), but to a lesser extent (Fig. 2e, Extended Data Table 1a). In the co-crystal structure of Cpf1 and crRNA, the targeting region of crRNA was not resolved, indicating that Cpf1 does not bind and stabilize this part of crRNA²¹. This is in contrast to Cas9, which makes extensive contact with the guide portion of the RNA, possibly explaining its longer seed region^{22,23}. Together with the recent Cpf1 characterization¹⁶, our results indicate that there may be additional factors influencing the specificity, such as the base content of the target sequence. The results highlight similarities between Cpf1 and Cas9 (refs 24, 25), which first recognizes the PAM and subsequently probes the crRNA complementary to the target DNA. Mismatches around the target site might disturb correct positioning of the catalytic residues and therefore reduce cleavage activity.



Aligning the two predicted protospacer sequences of the *F. novicida* U112 type V-A CRISPR-Cas revealed a conserved 5'-TTA-3' sequence located on the non-target strand upstream of the protospacer. To verify the potential PAM, protospacer 5 was cloned without its flanking region yielding a 5'-CTG-3' sequence. Both plasmids were cleaved equally well by Cpf1, indicating that the second position in this sequence is critical (Fig. 2d, Extended Data Fig. 7d). Mutagenesis of all three nucleotides followed by DNA cleavage analysis shows that Cpf1 recognizes a PAM, defined as 5'-YTN-3', upstream of the crRNA-complementary DNA sequence on the non-target strand. This result expands on the already reported 5'-TTN-3' PAM¹⁶. To analyse strand specificity of PAM recognition, we designed oligonucleotide substrates with either AAN or TTN on both strands. These substrates were not cleaved by Cpf1, indicating that the PAM needs to be double-stranded and is probably recognized on both strands (Fig. 2d, lower panel).

We next investigated the metal ion dependency of DNA cleavage by Cpf1. Notably, we observed that in addition to Mg²⁺ and Mn²⁺, which were shown to mediate activity in Cas9 (ref. 15), Cpf1 also cleaves DNA in the presence of Ca²⁺ (Extended Data Fig. 5b, Extended Data Table 1b). To investigate potential differences in cleavage with Mg²⁺ or Ca²⁺, we carried out DNA cleavage reactions in the presence of either of these ions (Fig. 2, Extended Data Fig. 7). In Cas9, two active motifs, HNH and RuvC, are responsible for cleavage of the target and non-target strand, respectively¹⁵. The HNH motif of Cas9 from *Neisseria meningitidis* is Ca²⁺-dependent²⁶. If there were two active sites in Cpf1, each coordinating one of the metal ions and cleaving one of the DNA strands, we would expect a difference in cleavage of target and non-target strands depending on the ion used. In contrast, we did not observe differences in the efficiency of target or non-target strand cleavage by Cpf1 in the presence of Ca²⁺ or Mg²⁺ (Fig. 2b, Extended Data Fig. 7b). This finding indicates the presence of only one catalytic motif in Cpf1 that is responsible for cleaving both DNA strands and can coordinate Mg²⁺ as well as Ca²⁺ ions.

Figure 2 | Cpf1 cleaves target DNA specifically at the 5'-YTN-3' PAM-distal end to generate 5-nt 5' overhangs in the presence of Ca²⁺. **a**, **b**, Cpf1-mediated target plasmid DNA cleavage (**a**) and Cpf1-mediated oligonucleotide duplex cleavage (**b**), dependent on the crRNA containing spacer 4 or 5 (crRNA-sp4 or crRNA-sp5), in the absence or presence of Ca²⁺. **c**, Schematic representation of the protospacer 5 sequence in the DNA (top), and the structure of crRNA-sp5 used in **a**, **b**, **d** and **e** (bottom). Cleavage sites corresponding to fragments obtained in **b** and confirmed by sequencing (Extended Data Fig. 7) are indicated by blue triangles. The PAM is marked in grey. **d**, Plasmid DNA containing the PAMs 1–6, or 5'-radiolabelled double-stranded oligonucleotide containing PAMs 1, 7–9 were cleaved by Cpf1 in the presence of 10 mM CaCl₂ (upper and lower panel, respectively). **e**, Plasmids containing protospacer 5 and single or quadruple mismatches (mut_1–4 and mut_19–22) along the target strand were tested for cleavage by Cpf1 programmed with crRNA-sp5 in the presence of 10 mM MgCl₂. Quantification of three independent experiments are shown in Extended Data Table 1a. li, linear; sc, supercoiled; M, 1 kb ladder. Data in **a**, **b**, **d** and **e** are representatives of at least three independent experiments.

Our experiments show for the first time that Cpf1 exhibits dual (RNA and DNA) cleavage activity. To determine the respective cleavage motifs, we performed mutagenesis of conserved residues along the Cpf1 amino acid sequence (Supplementary Fig. 2). Alanine substitution of residues H843, K852, K869 and F873 had no effect on DNA cleavage activity (Fig. 3a, upper panel), but resulted in decreased *in vitro* RNA cleavage activity (Fig. 3a, middle panel). To further confirm the involvement of these residues in RNA processing *in vivo*, a heterologous *E. coli* assay co-expressing pre-crRNA (repeat-spacer-repeat) and Cpf1, or a variant thereof, was established. Northern blot analysis was performed with total RNA extracted following induced expression of the Cpf1 variant (Fig. 3a, lower panel, Extended Data Fig. 3b). Pre-crRNA is more abundant in the presence of Cpf1, indicating possible protection from degradation by Cpf1. Expression of wild-type Cpf1 results in the production of a distinct band of around 65 nt, which corresponds to a mature crRNA formed by two cleavage events within the repeats. In the presence of Cpf1(H843A), this band is absent; however, two additional longer RNAs appear due to changed processing by this mutant, as observed *in vitro* (Fig. 3a, middle panel). Mutants K852A and K869A also resulted in the production of the 65-nt fragment, but with less intensity compared to the wild type, and two additional products with longer sizes. *In vitro*, these mutants have almost no RNA processing activity. RNA-binding experiments with Cpf1(K852A) and Cpf1(K869A) (Extended Data Fig. 8b) indicated a slightly higher affinity for RNA than wild-type Cpf1, which may explain the cleavage products observed *in vivo*. The residual activity of these Cpf1 mutants produces processed RNA, which is likely to be bound tighter to the protein and therefore better protected from degradation. Cpf1(F873A) had reduced RNA cleavage activity *in vitro*, which could not be detected *in vivo*. Mutation of the aforementioned residues did not negatively affect RNA binding (Extended Data Fig. 8b), indicating that the identified residues of Cpf1 are potentially responsible for RNA cleavage. Analysis of the co-crystal structure of *Lachnospiraceae* bacterium Cpf1

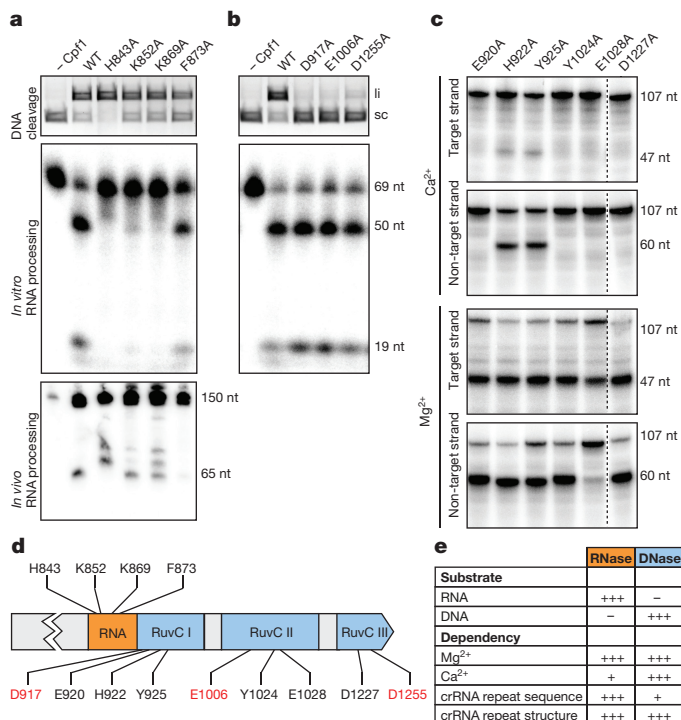


Figure 3 | Cpf1 contains active centres for RNA and DNA cleavage. **a**, RNase motif mutants were tested for DNA plasmid cleavage activity (agarose gel, upper panel), *in vitro* pre-crRNA cleavage activity (denaturing polyacrylamide gel, middle panel) and *in vivo* pre-crRNA processing activity (northern blot, lower panel). *In vitro* cleavage was performed in the presence of 10 mM MgCl₂. **b**, DNase motif mutants were tested for plasmid DNA cleavage activity (agarose gel, upper panel) and *in vitro* pre-crRNA cleavage activity (denaturing polyacrylamide gel, lower panel). **c**, Additional RuvC motif mutants were tested for DNA cleavage of double-stranded oligonucleotide substrates in 10 mM CaCl₂ (upper two panels) or MgCl₂ (lower two panels). Target or non-target strand was 5' radiolabelled before annealing to the non-labelled complementary strand. **d**, Schematic representation of Cpf1 amino acid sequence (N terminus not shown for clearer visualization) with the active domains for RNA and DNA cleavage highlighted in orange and blue, respectively. Mutated amino acids are indicated with the DNase motif shown in red. **e**, Summary of recognized substrates, metal ion dependency and crRNA requirements for both RNase and DNase motifs of Cpf1. —, no activity; + residual activity; +++ full activity. Data in **a–c** are representatives of at least three independent experiments.

revealed that the identified residues are located in close proximity to the 5' of the processed crRNA²¹.

Mutagenesis of D917, E1006 and D1255 in the split RuvC motif resulted in loss of DNA cleavage activity¹⁶ (Fig. 3b, upper panel), but did not influence the RNA processing activity of Cpf1 (Fig. 3b, lower panel), nor did it affect binding affinity to the DNA target (Extended Data Fig. 8c).

While screening for active site residues, we observed differences in DNA cleavage for some mutants depending on the metal ion present. Mutants E920A, Y1024A and D1227A showed no DNA cleavage in the presence of Ca²⁺, but wild-type activity when Mg²⁺ was present (Fig. 3c). These residues are located in close proximity to the three identified catalytic residues and may be responsible for coordination of the Ca²⁺ ion. Mutating residue E1028 also led to loss of Ca²⁺-promoted DNA cleavage and additionally reduced cleavage of the non-target strand in the presence of Mg²⁺, indicative of its involvement in non-target strand cleavage. In contrast, mutation of residues H922 and Y925 resulted in markedly reduced cleavage of the target strand in the presence of Ca²⁺, whereas these mutants showed wild-type levels of DNA cleavage activity in the presence of Mg²⁺.

These findings suggest that H922 and Y925 are involved in Ca²⁺ coordination and target-strand cleavage.

We show that two aspartates (D917, D1255) and one glutamate (E1006) form the catalytic site, which is in good agreement with the recent characterization of Cpf1 and other RuvC/RNaseH motifs¹⁶. These kind of catalytic motifs generally use a two-metal-ion mechanism for DNA cleavage, as shown for Cas9 from *Streptococcus pyogenes*²³. Enzymes with a two-metal-ion mechanism have more specificity for metal ions, Mg²⁺ in particular²⁷. In contrast, enzymes using a one-metal-ion mechanism for cleavage (for example, HNH nucleases) are more flexible in their specificity for metal ions. For example, KpnI cleaves DNA with high fidelity in the presence of Ca²⁺, but less specifically in the presence of Mg²⁺ (ref. 28). As mentioned before, the HNH motif of Cas9 from *N. meningitidis* is active in the presence of Ca²⁺ (ref. 26). In addition to the identified RNA processing activity of Cpf1, this enzyme may also represent a new type of DNA nuclease using two-metal-ion catalysis, with the ability to utilize Mg²⁺ or Ca²⁺ ions. The physiological relevance of Cpf1 using both ions for DNA cleavage remains undetermined and requires further investigation.

In summary, Cpf1 is an enzyme with two separate catalytic moieties that cleave RNA or DNA (Fig. 3d). The RNase motif is specific for the ribose and unable to cleave DNA. This specificity can be explained by specific interactions of Cpf1 to 2'-OH groups of crRNA²¹. The DNase motif shows cleavage activity only against double-stranded and single-stranded target DNA, but no activity against single-stranded RNA, double-stranded RNA or RNA–DNA heteroduplexes (Fig. 3e, Extended Data Fig. 9). There are other nucleases reported to have certain promiscuity towards RNA and DNA cleavage activity, but one of the two activities is usually highly unspecific^{29,30}. To our knowledge, Cpf1 is the first enzyme with two specificities, cleaving RNA in a sequence- and structure-dependent manner, and also performing DNA cleavage in the presence of the RNA that is produced in the first reaction. In the context of CRISPR immunity, type V-A appears to be the most minimalistic system described thus far, using only one enzyme, Cpf1, to process pre-crRNA and then using this RNA to specifically target and cut invading DNA. Evolution of one protein to perform these two specific reactions leads to a more effective mechanism, and also makes this system ideal for horizontal gene transfer. Finally, this mechanism opens new avenues for sequence-specific genome engineering, silencing and facilitates multiplexing.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 20 November 2015; accepted 30 March 2016.

Published online 20 April 2016.

- Marraffini, L. A. CRISPR-Cas immunity in prokaryotes. *Nature* **526**, 55–61 (2015).
- Carte, J., Wang, R., Li, H., Terns, R. M. & Terns, M. P. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes Dev.* **22**, 3489–3496 (2008).
- Ebihara, A. et al. Crystal structure of hypothetical protein TTHB192 from *Thermus thermophilus* HB8 reveals a new protein family with an RNA recognition motif-like domain. *Protein Sci.* **15**, 1494–1499 (2006).
- Charpentier, E., Richter, H., van der Oost, J. & White, M. F. Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. *FEMS Microbiol. Rev.* **39**, 428–441 (2015).
- Nam, K. H. et al. Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/*Dvulg* CRISPR-Cas system. *Structure* **20**, 1574–1584 (2012).
- Jore, M. M. et al. Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nature Struct. Mol. Biol.* **18**, 529–536 (2011).
- Mulepati, S., Heroux, A. & Bailey, S. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. *Science* **345**, 1479–1484 (2014).
- Plagens, A., Richter, H., Charpentier, E. & Randau, L. DNA and RNA interference mechanisms by CRISPR-Cas surveillance complexes. *FEMS Microbiol. Rev.* **39**, 442–463 (2015).
- van der Oost, J., Westra, E. R., Jackson, R. N. & Wiedenheft, B. Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nature Rev. Microbiol.* **12**, 479–492 (2014).

10. Zhang, J. *et al.* Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. *Mol. Cell* **45**, 303–313 (2012).
11. Staals, R. H. *et al.* RNA targeting by the type III-A CRISPR-Cas Csm complex of *Thermus thermophilus*. *Mol. Cell* **56**, 518–530 (2014).
12. Samai, P. *et al.* Co-transcriptional DNA and RNA cleavage during type III CRISPR-Cas immunity. *Cell* **161**, 1164–1174 (2015).
13. Chylinski, K., Le Rhun, A. & Charpentier, E. The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. *RNA Biol.* **10**, 726–737 (2013).
14. Deltcheva, E. *et al.* CRISPR RNA maturation by *trans*-encoded small RNA and host factor RNase III. *Nature* **471**, 602–607 (2011).
15. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
16. Zetsche, B. *et al.* Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* **163**, 759–771 (2015).
17. Schunder, E., Rydzewski, K., Grunow, R. & Heuner, K. First indication for a functional CRISPR/Cas system in *Francisella tularensis*. *Int. J. Med. Microbiol.* **303**, 51–60 (2013).
18. Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nature Rev. Microbiol.* **13**, 722–736 (2015).
19. Shmakov, S. *et al.* Discovery and functional characterization of diverse class 2 CRISPR-Cas systems. *Mol. Cell* **60**, 385–397 (2015).
20. Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K. & Doudna, J. A. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* **329**, 1355–1358 (2010).
21. Dong, D. *et al.* The crystal structure of Cpf1 in complex with CRISPR RNA. *Nature* <http://dx.doi.org/10.1038/nature17944> (20 April 2016).
22. Jiang, F., Zhou, K., Ma, L., Gressel, S. & Doudna, J. A. A Cas9-guide RNA complex preorganized for target DNA recognition. *Science* **348**, 1477–1481 (2015).
23. Nishimasu, H. *et al.* Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935–949 (2014).
24. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62–67 (2014).
25. Szczelkun, M. D. *et al.* Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proc. Natl Acad. Sci. USA* **111**, 9798–9803 (2014).
26. Zhang, Y., Rajan, R., Seifert, H. S., Mondragon, A. & Sontheimer, E. J. DNase H activity of *Neisseria meningitidis* Cas9. *Mol. Cell* **60**, 242–255 (2015).
27. Yang, W. An equivalent metal ion in one- and two-metal-ion catalysis. *Nature Struct. Mol. Biol.* **15**, 1228–1231 (2008).
28. Vasu, K. *et al.* Increasing cleavage specificity and activity of restriction endonuclease *KpnI*. *Nucleic Acids Res.* **41**, 9812–9824 (2013).
29. Nam, K. H. *et al.* Double-stranded endonuclease activity in *Bacillus halodurans* clustered regularly interspaced short palindromic repeats (CRISPR)-associated Cas2 protein. *J. Biol. Chem.* **287**, 35943–35952 (2012).
30. Punetha, A., Sivathanu, R. & Anand, B. Active site plasticity enables metal-dependent tuning of Cas5d nuclease activity in CRISPR-Cas type I-C system. *Nucleic Acids Res.* **42**, 3846–3856 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank K. Schmidt, F. Hille and A. Escalera Maurer for technical help. This work was funded by the Alexander von Humboldt Foundation (AvH Professorship), the German Federal Ministry for Education and Research, the Helmholtz Association, the German Research Foundation, the Max Planck Society, the Göran Gustafsson Foundation (Göran Gustafsson Prize from the Royal Swedish Academy of Sciences), the Swedish Research Council and Umeå University (all to E.C.), and the Helmholtz Postdoc Programme (to H.R.).

Author Contributions I.F. and H.R. conducted the biochemical characterization of the DNase and RNase activities, M.B. performed binding studies and seed sequence characterization and A.L.R. performed and analysed RNA sequencing. I.F., H.R. and E.C. designed the research. I.F., H.R., M.B., A.L.R. and E.C. analysed and interpreted the data; I.F., H.R., and E.C. wrote the paper, which M.B. and A.L.R. commented on.

Author Information RNA sequencing data have been deposited at NCBI under accession number SRP071054. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.C. (charpentier@mpiib-berlin.mpg.de).

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Small RNA sequencing. Small RNA sequencing data of *Francisella novicida* U112 (Supplementary Table 1a) used in this study were obtained previously¹³. Briefly, a cDNA library of RNAs (treated with tobacco acid pyrophosphatase) of *F. novicida* U112 grown to mid-logarithmic phase was prepared using the ScriptMiner Small RNA-Seq Library Preparation Kit (Multiplex, Illumina compatible) and sequenced at the Campus Science Support Facilities GmbH (CSF) Next Generation Sequencing (NGS) Unit of the Vienna Biocentre. After adaptor removal and quality trimming, the reads were mapped to the *F. novicida* U112 genome (GenBank: NC_008601, 48205 mapped reads) using Bowtie. The read coverage was calculated using BEDTools (version 2.15.0.)³¹ and a normalized wiggle file was created and visualized using the Integrative Genomics Viewer^{32,33}.

Production and purification of recombinant Cpf1. The *cpf1* (FTN_1397) gene was amplified from genomic DNA of *F. novicida* U112 and cloned into the expression vector pET-16b to facilitate expression of Cpf1 with an N-terminal 6× His-tag (Supplementary Table 1b, c). For the production of the protein in *E. coli* (NiCo21 (DE3)), the cells containing the overexpression plasmid were grown at 37°C to reach an optical density (OD)₆₀₀ of 0.6–0.8. Expression was induced by addition of 0.5 mM isopropylthio-β-D-galactoside (IPTG) and the cultures were further incubated overnight at 18°C. After collection, the cell pellet was resuspended in lysis buffer (20 mM HEPES (pH 7.5), 500 mM KCl, 25 mM imidazole, 0.1% Triton X-100) followed by 6 min of sonication (0.5 s pulses) for cell disruption. The lysate was cleared by centrifugation (47,800g, 30 min, 4°C) and the supernatant was applied to Ni²⁺-NTA-Sepharose resin in a drop column. After washing steps with 10 ml of lysis buffer followed by 10 ml wash buffer (20 mM HEPES (pH 7.5), 300 mM KCl, 25 mM imidazole), the protein was eluted with elution buffer (20 mM HEPES (pH 7.5), 150 mM KCl, 250 mM imidazole, 0.1 mM DTT, 1 mM EDTA). The eluates were analysed by SDS–PAGE followed by Coomassie blue staining. Fractions containing Cpf1 were pooled for cation-exchange chromatography (HiTrap Heparin; GE-Healthcare) using a FPLC Äkta-Purification system (GE-Healthcare) and Cpf1 was eluted with a linear gradient of 100–1000 mM KCl. Peak fractions were analysed by SDS–PAGE and Coomassie blue staining. Cpf1-containing fractions were pooled and directly applied to an equilibrated (20 mM HEPES (pH 7.5), 150 mM KCl) pregrade Superdex 200 size-exclusion column (GE-Healthcare) and purified via fast protein liquid chromatography (FPLC), followed by analysis by SDS–PAGE and Coomassie blue staining. Molecular weight calibration of the column was performed using molecular weight markers, as described in the manufacturer's protocol (Kit for Molecular Weights, Sigma-Aldrich). The protein was dialysed against dialysis buffer (20 mM HEPES (pH 7.5), 150 mM KCl, 50% glycerol) and stored at –20°C until use.

Site-directed mutagenesis of Cpf1. Oligonucleotides for the site-directed mutation of Cpf1 (Supplementary Table 1c) were designed using the QuickChange Primer Design tool of Agilent and produced by Sigma-Aldrich. A series of PCRs was performed to obtain the desired mutation. Briefly, the overexpression vector containing wild-type *cpf1* was amplified in two reactions with either the forward or reverse QuickChange primer. After an initial amplification, the two reactions were mixed and a second PCR was performed. Following PCR, the template plasmid was degraded with DpnI (3 h, 37°C) and introduced by transformation into chemically competent DH5α cells. Plasmids were prepared using a plasmid Miniprep kit (Qiagen) according to the manufacturer's instructions. Successful mutagenesis was confirmed by sequencing analysis of the plasmids (SeqLab).

Generation of RNAs used in this study. The sRNAs tested in this study were generated by *in vitro* transcription using the AmpliScribe T7-Flash kit (Biozym) according to the manufacturer's protocol. In brief, oligonucleotides containing the desired sequence (Supplementary Table 1c) and a T7-promoter sequence were hybridized to an oligonucleotide containing the complementary T7-promoter sequence. The hybridization product was then used as a template for the transcription reaction according to the AmpliScribe T7-Flash kit (Biozym). To obtain internally labelled RNAs, [α-³²P]ATP (5000 Ci mmol^{–1}, Hartman Analytic) was added to the *in vitro* transcription reaction³⁴. In order to generate end-labelled RNAs, the unlabelled transcripts were dephosphorylated with Fast-AP phosphatase (Fermentas) for 30 min at 37°C followed by a purification using Illustra Microspin G-25 columns (GE-Healthcare). The dephosphorylated RNAs were then labelled using T4 polynucleotide kinase (Fermentas) and [γ-³²P]ATP (5000 Ci mmol^{–1}) according to the manufacturer's instructions, and separated using denaturing polyacrylamide gel electrophoresis (8 M urea; 1× TBE; 10% polyacrylamide). Subsequent to short exposure to an autoradiography screen (for radioactively labelled RNAs) or ethidium bromide (EtBr) staining (for unlabelled RNAs), the respective bands of the RNAs were excised. Elution of the RNAs was achieved by

incubation of the gel pieces in 500 μl RNA elution buffer (250 mM NaOAc; 20 mM Tris-HCl (pH 7.5); 1 mM EDTA (pH 8.0); 0.25% SDS) and overnight incubation on ice. Following elution, RNA was precipitated with 2 vol ice-cold ethanol (100% EtOH) and 1/100 glycogen for 1 h at –20°C. After washing with 70% EtOH, the air-dried pellets were resuspended in H₂O.

In vitro RNA cleavage assay. RNA cleavage assays using indicated concentrations of Cpf1 and various RNA substrates were conducted in KGB buffer³⁵ (100 mM potassium glutamate, 25 mM Tris-acetate (pH 7.5), 500 μM 2-mercaptoethanol, 10 μg ml^{–1} BSA) supplemented with 10 mM MgCl₂ at 37°C in a final volume of 10 μl. If not indicated otherwise, the reaction was stopped after 10 min by the addition of 2 μl proteinase K (20 mg ml^{–1}) following 10 min incubation at 37°C to achieve protein degradation. After adding 2× loading dye (10 M urea, 1.5 mM EDTA (pH 8.0)), the samples were loaded on 12% denaturing polyacrylamide gels run in 1× TBE for 3 h at 12.5 V cm^{–1}. For the sequencing gels, the samples were precipitated before loading on 10% denaturing polyacrylamide gels. The gel electrophoresis was carried out at 40 W for 3.5 h. Visualization was achieved by phosphorimaging (Typhoon FLA 9000 Fuji). For RNA size determination, a 5'-end-labelled 69-nt long transcript consisting of a short form of pre-crRNA (repeat-spacer 5, full-length) was subjected to alkaline hydrolysis generating a single nucleotide resolution ladder and to RNase T1-specific cleavage. Each individual experiment was performed in three replicates.

In vivo RNA processing. To investigate *in vivo* RNA processing by Cpf1, a heterologous system was designed in *E. coli*. A DNA fragment encoding a pre-crRNA containing a repeat-spacer-repeat structure under the control of a T7-promoter and T7-terminator was synthesized by Integrated DNA Technologies and cloned into pACYC184 using HindIII and EagI yielding pEC1690. *E. coli* BL21(DE3) was co-transformed with this plasmid and the overexpression vector of wild-type or mutant Cpf1. The empty expression vector pET-16b served as a negative control. The bacterial cells were grown in the presence or absence of 0.1 mM IPTG at 37°C to reach early exponential phase (OD₆₀₀ = 0.4). RNA was extracted using TRIzol (Sigma-Aldrich) according to the manufacturer's protocol followed by northern blot analysis as described previously^{36–38}. In brief, RNA was separated on denaturing 10% polyacrylamide gels (8 M urea, 1× TBE) and transferred by semi-dry blotting on a nylon membrane (Hybond TM N+, GE Healthcare). Chemical crosslinking was performed for 1 h at 60°C with 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide hydrochloride. Oligonucleotides were radioactively labelled with [γ-³²P]ATP (5000 Ci mmol^{–1}) and T4 polynucleotide kinase (Fermentas) as described above and purified using Illustra Microspin G-25 columns (GE Healthcare). The hybridization of the probe against the spacer in the pre-crRNA (Supplementary Table 1c) was performed in Rapid-hyb buffer (GE-Healthcare) by incubation overnight at 42°C. The radioactive signal was visualized using phosphorimaging. Each individual experiment was performed in three replicates.

Generation of DNA substrates. To determine the target cleavage site of Cpf1, spacer sequences of the *F. novicida* U112 type V-A CRISPR array were analysed by BLAST³⁹. Potential targets for spacer 4 and spacer 5 were identified in *F. novicida* 3523, located in the intergenic region between coding sequence AEE26308.1 and AEE26307.1, and in AEE26301.1, respectively. Target protospacer containing a sequence complementary to spacer 5 including 42 bp up- and downstream sequences was synthesized as oligonucleotides containing HindIII overhangs. Following hybridization of the oligonucleotides, the fragments were cloned into pUC19 using HindIII yielding plasmid pEC1664 (protospacer 5 and flanking region). The same protospacer sequence without flanking regions was cloned into pUC19, yielding pEC1688 (protospacer 5). In order to identify the PAM, mutagenesis was performed by applying the described protocol for site-directed mutagenesis on pEC1688. Plasmid preparation was done using Miniprep kit (Qiagen) according to the manufacturer's instructions and DNA integrity was confirmed by sequencing analysis (SeqLab). Oligonucleotides containing the protospacer (Supplementary Table 1c) were ordered at Sigma and hybridized before radioactive labelling. Alternatively, a single-stranded oligonucleotide was labelled and hybridized with the complementary non-labelled oligonucleotide. 5'-end-labelling reactions were performed using [γ-³²P]ATP (5000 Ci mmol^{–1}) and T4 polynucleotide kinase (Fermentas) according to the manufacturer's instructions. The labelled oligonucleotides were purified using Illustra Microspin G-25 columns (GE healthcare).

In vitro DNA cleavage assay. Plasmid DNA cleavage assays were performed by pre-incubating 100 nM Cpf1 with 200 nM RNA in KGB buffer supplemented with either 10 mM MgCl₂ or 10 mM CaCl₂ for 15 min at 37°C. Plasmid DNA (10 nM) was added to the reaction to yield a final volume of 10 μl and further incubated for 1 h at 37°C. Reactions were stopped by the addition of 1 μl proteinase K (20 mg ml^{–1}) and 5 min incubation at 37°C. Before separation of the reaction, 3 μl of 5× DNA loading buffer (250 mM EDTA, 1.2% SDS, 25% glycerol, 0.01% bromophenol blue) were added and the samples were loaded on 0.8% agarose gels

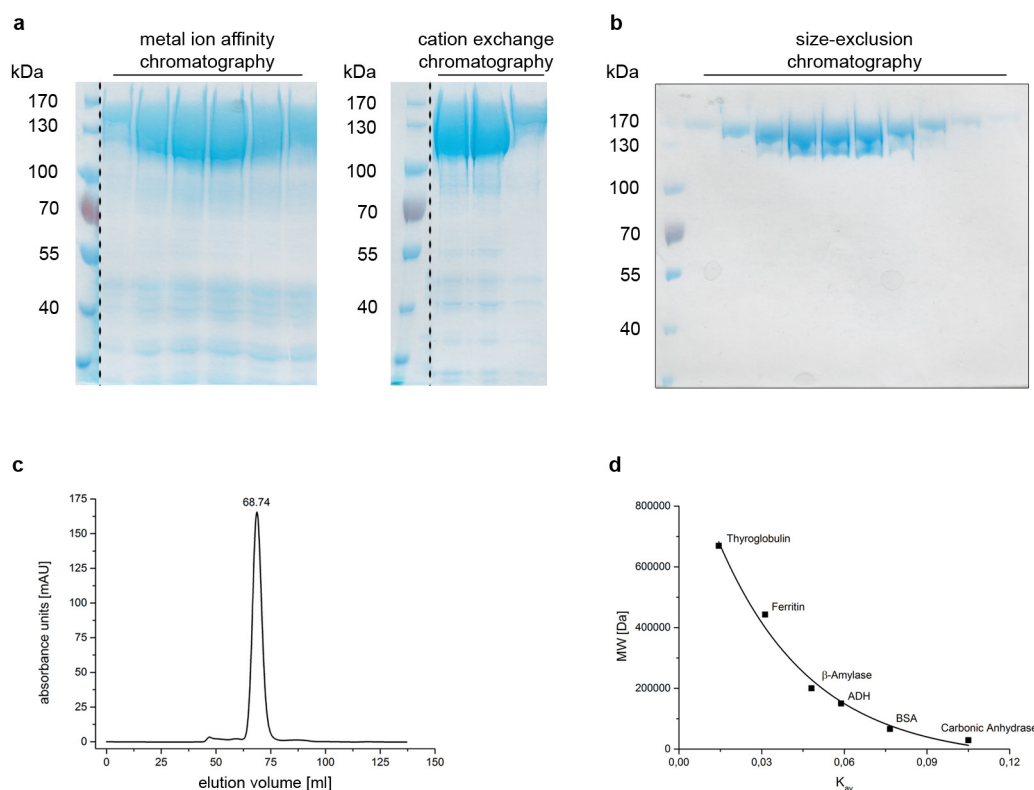
(1× TAE buffer). Cleavage products were visualized by EtBr staining. In cleavage assays using radioactively labelled substrates, 5 nM of 5'-labelled double-stranded oligonucleotides were added to the pre-formed complex of CpfI and RNA, and incubated at 37°C for 1 h. After proteinase K treatment, 10 µl of 2× denaturing loading buffer (95% formamide, 0.025% SDS, 0.5 mM EDTA, 0.025% bromophenol blue) were added. Oligonucleotides of the size of the expected cleavage products were 5'-radiolabelled as described above and mixed with an equal volume of 2× denaturing loading buffer to serve as size markers. After 5 min incubation at 95°C, the samples were loaded on 12% denaturing polyacrylamide gels and run in 1× TBE for 70 min at 14 V cm⁻¹. Cleavage was visualized using phosphorimaging. Each individual experiment was performed in three replicates.

Electrophoretic mobility shift assays. Substrates for electrophoretic mobility shift assays (EMSAs) were generated as described above. For DNA binding reactions, CpfI was pre-incubated in binding buffer (20 mM Tris-HCl (pH 7.4), 100 mM KCl, 1 mM DTT, 5% glycerol) containing two molar excess of crRNA. After 15 min at 37°C, 1 nM labelled DNA substrate was added. The reaction was then carried out at 37°C for 1 h before the samples were loaded on a native 5% polyacrylamide gel running at 10 V cm⁻¹ for 50 min in 0.5× TBE to separate protein-DNA complexes from unbound DNA. For RNA binding reactions, the crRNA was dephosphorylated using Fast AP (Fermentas) and 5'-radiolabelled with [γ -³²P] ATP (5000 Ci mmol⁻¹) and T4 polynucleotide kinase (Fermentas) according to the manufacturer's instructions. A total of 0.5 nM radiolabelled RNA were incubated with CpfI in binding buffer (20 mM Tris (pH 7.5), 150 mM KCl, 10 mM CaCl₂, 1 mM DTT, 5% glycerol, 0.01% Triton X-100, 10 µg ml⁻¹ BSA) for 1 h at 37°C and loaded on 4% native polyacrylamide gels running at 10 V cm⁻¹ for 30 min in 0.5× TBE. The gels were exposed on an autoradiography film overnight and visualized by phosphorimaging. Fractions of bound and unbound nucleic acids were determined densitometrically and the percentage of bound nucleic acid was plotted against the protein concentrations. The dissociation constant, K_d , was determined using a nonlinear regression analysis.

Multiple sequence alignment of CpfI orthologues. CpfI orthologous sequences were derived by BLAST³⁹ search of the NCBI database using CpfI of *F. novicida*

U112 as a query. A multiple sequence alignment of 52 orthologous sequences was generated using MUSCLE⁴⁰. The alignment of nine of the sequences was visualized with Jalview⁴¹.

31. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
32. Robinson, J. T. et al. Integrative genomics viewer. *Nature Biotechnol.* **29**, 24–26 (2011).
33. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
34. Sampson, J. R. & Uhlenbeck, O. C. Biochemical and physical characterization of an unmodified yeast phenylalanine transfer RNA transcribed *in vitro*. *Proc. Natl Acad. Sci. USA* **85**, 1033–1037 (1988).
35. McClelland, M., Hanish, J., Nelson, M. & Patel, Y. KGB: a single buffer for all restriction endonucleases. *Nucleic Acids Res.* **16**, 364 (1988).
36. Herbert, S., Barry, P. & Novick, R. P. Subinhibitory clindamycin differentially inhibits transcription of exoprotein genes in *Staphylococcus aureus*. *Infect. Immun.* **69**, 2996–3003 (2001).
37. Urban, J. H. & Vogel, J. Translational control and target recognition by *Escherichia coli* small RNAs *in vivo*. *Nucleic Acids Res.* **35**, 1018–1037 (2007).
38. Pall, G. S. & Hamilton, A. J. Improved northern blot method for enhanced detection of small RNA. *Nature Protocols* **3**, 1077–1084 (2008).
39. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
40. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
41. Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
42. Robinson, J. T. et al. Integrative genomics viewer. *Nature Biotechnol.* **29**, 24–26 (2011).
43. Gruber, A. R., Lorenz, R., Bernhart, S. H., Neubock, R. & Hofacker, I. L. The Vienna RNA websuite. *Nucleic Acids Res.* **36**, 3429–3431 (2008).
44. Darty, K., Denise, A. & Ponty, Y. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25**, 1974–1975 (2009).



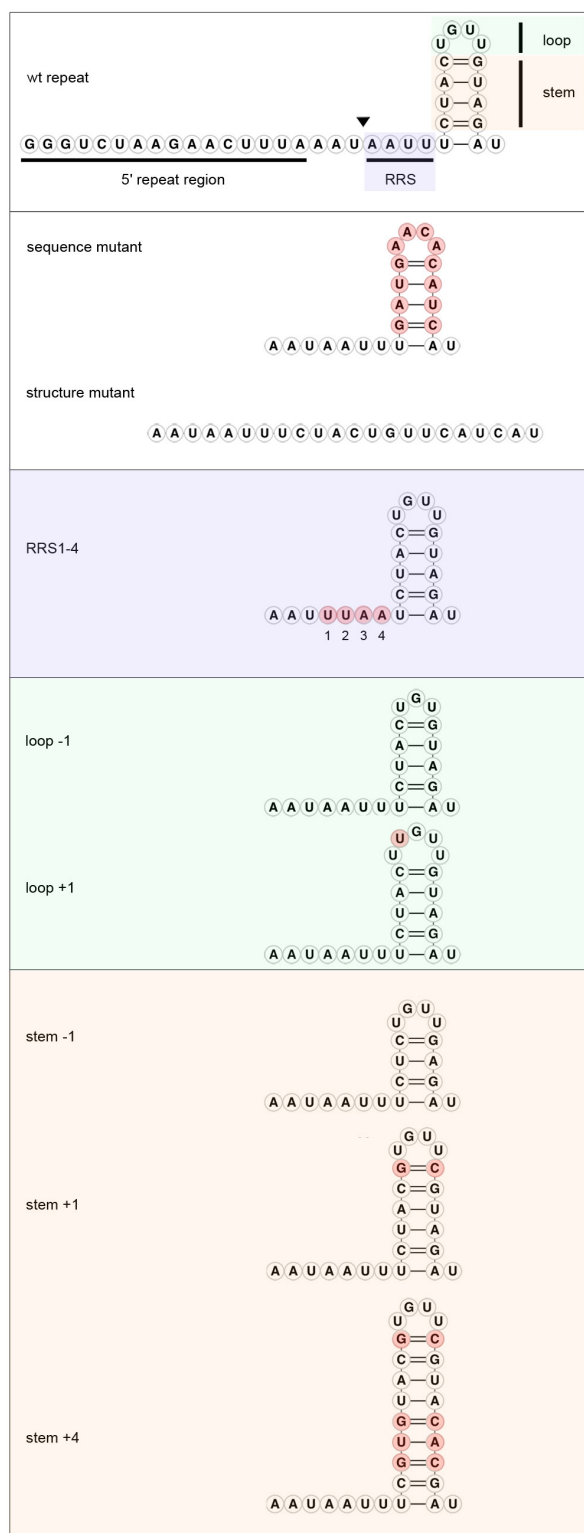
Extended Data Figure 2 | Wild-type Cpf1 purifies as a monomer in solution. Recombinant Cpf1 of *F. novicida* U112 purified via affinity and cation-exchange chromatography was applied to a Superdex 200 size-exclusion column. **a**, SDS-PAGE of protein eluates obtained by nickel-affinity purification (left panel), which were further purified by cation-exchange chromatography (right panel). **b**, Protein samples obtained by size-exclusion chromatography were separated by SDS-PAGE (8% polyacrylamide) and visualized with Coomassie staining. **c**, Elution profile of the size-exclusion chromatography of wild-type Cpf1. The partition coefficient K_{av} for Cpf1 was calculated as 0.0538 by using the equation $K_{av} = (V_e - V_0)/(V_t - V_0)$, with V_e , elution volume; V_0 , void volume (elution volume of blue dextran, 45.171 ml) and V_t , geometric column

volume (482.5 ml). **d**, Calibration curve of proteins with known molecular weights (thyroglobulin (669 kDa), ferritin (443 kDa), β -amylase (200 kDa), alcohol dehydrogenase (ADH; 150 kDa), bovine serum albumin (BSA; 66 kDa), carbonic anhydrase (29 kDa); Molecular Weight Marker Kit, Sigma-Aldrich). The molecular weight of these proteins was plotted against their calculated K_{av} and fitted by exponential regression analysis. On the basis of the calculation, the K_{av} of Cpf1 results in a molecular weight of 187 kDa, indicating a monomeric form of Cpf1 in solution. Assuming that the protein does not adopt a perfect globular shape, this result is in accordance with the theoretical molecular weight of Cpf1 (153 kDa).

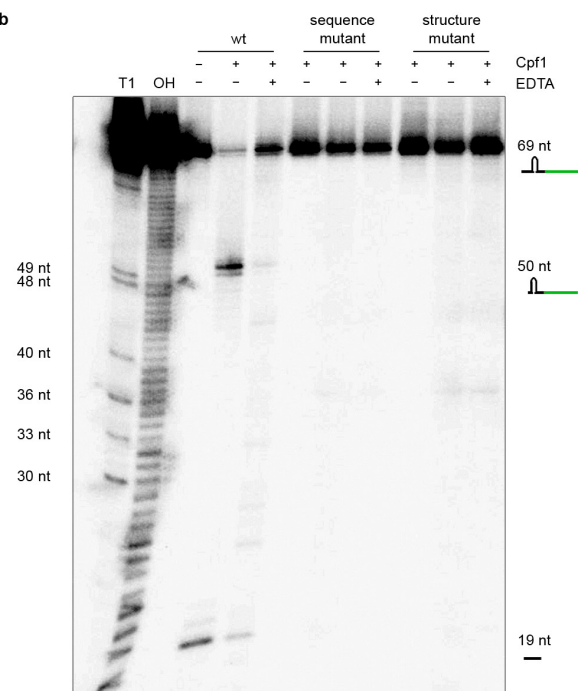
Extended Data Figure 3 | The endoribonucleolytic activity of Cpf1 is dependent on the presence of an intact repeat sequence. **a**, Cleavage assays were performed by incubating 100 nM of internally labelled RNA constructs corresponding to different repeat and spacer sequence variants of pre-crRNA-sp5 (pre-crRNA-containing spacer 5) with 1 μ M of Cpf1 for 30 min at 37 °C. The cleavage reaction was analysed by denaturing polyacrylamide gel electrophoresis and phosphorimaging. The cleavage products are represented schematically. The sequence compositions of the RNAs used as substrates are shown. RNA structures were generated with RNAfold⁴³ and visualized using VARNA⁴⁴ software. Cpf1 cleaved only the RNA templates containing a full-length repeat sequence. The substrate containing two repeats was cleaved twice resulting in more than two fragments, whereas cleavage of RNA with only one repeat resulted in two fragments, consistent with the determined cleavage site

(see Fig. 1). **b**, Northern blot analysis of total RNA extracted from *E. coli* co-transformed with a plasmid encoding pre-crRNA and either the empty vector or overexpression vectors encoding wild-type (wt) Cpf1 and variants. Cpf1 expression was induced (+) or not induced (–) with IPTG. The northern blot was probed against the spacer sequence of the tested pre-crRNA. In the absence of Cpf1 (empty vector or not induced), the amount of transcript is reduced compared to reactions with Cpf1 present, suggesting stabilization of pre-crRNA upon binding of Cpf1. Expression of Cpf1, Cpf1(K852A) and Cpf1(K869A) results in the production of a distinct processed transcript of 65 nt, whereas expression of Cpf1(H843A), Cpf1(K852A) or Cpf1(K869A) results in the production of additional higher transcripts. Expression of Cpf1(F873A) results in almost no detectable processed transcript.

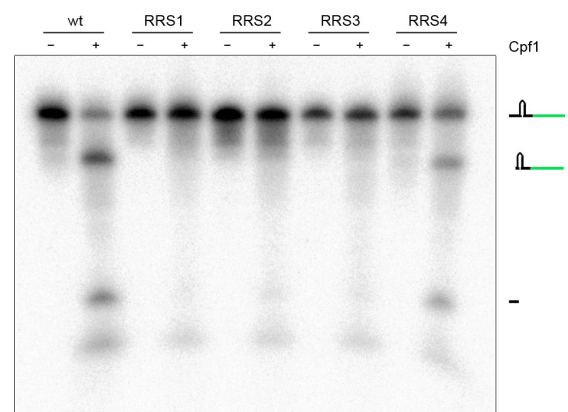
a



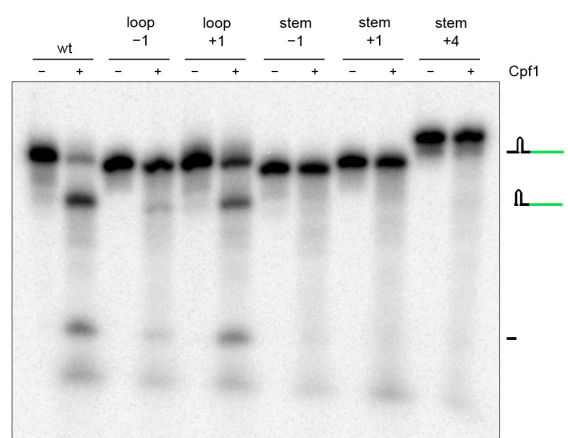
b



c



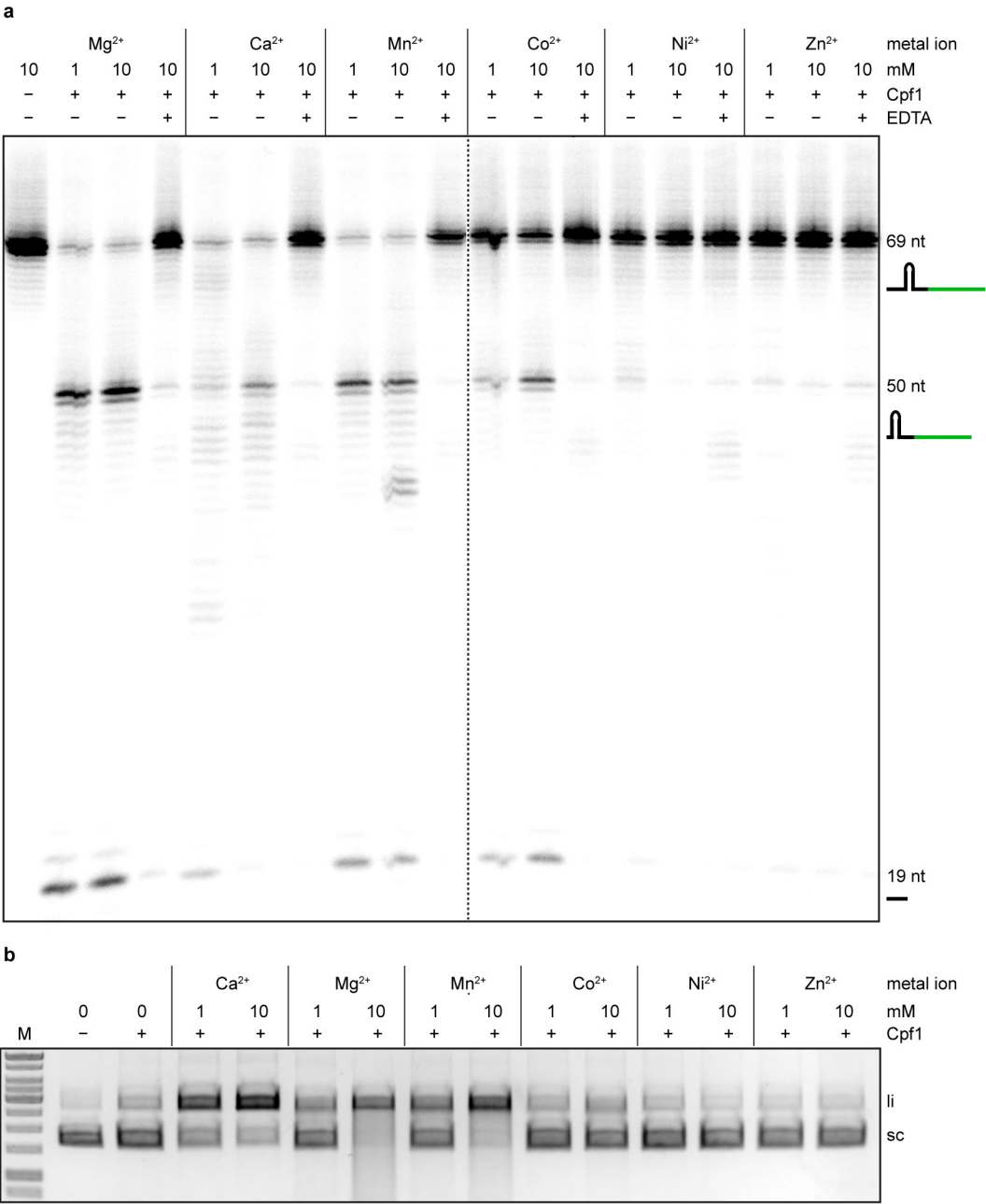
d



Extended Data Figure 4 | See next page for caption.

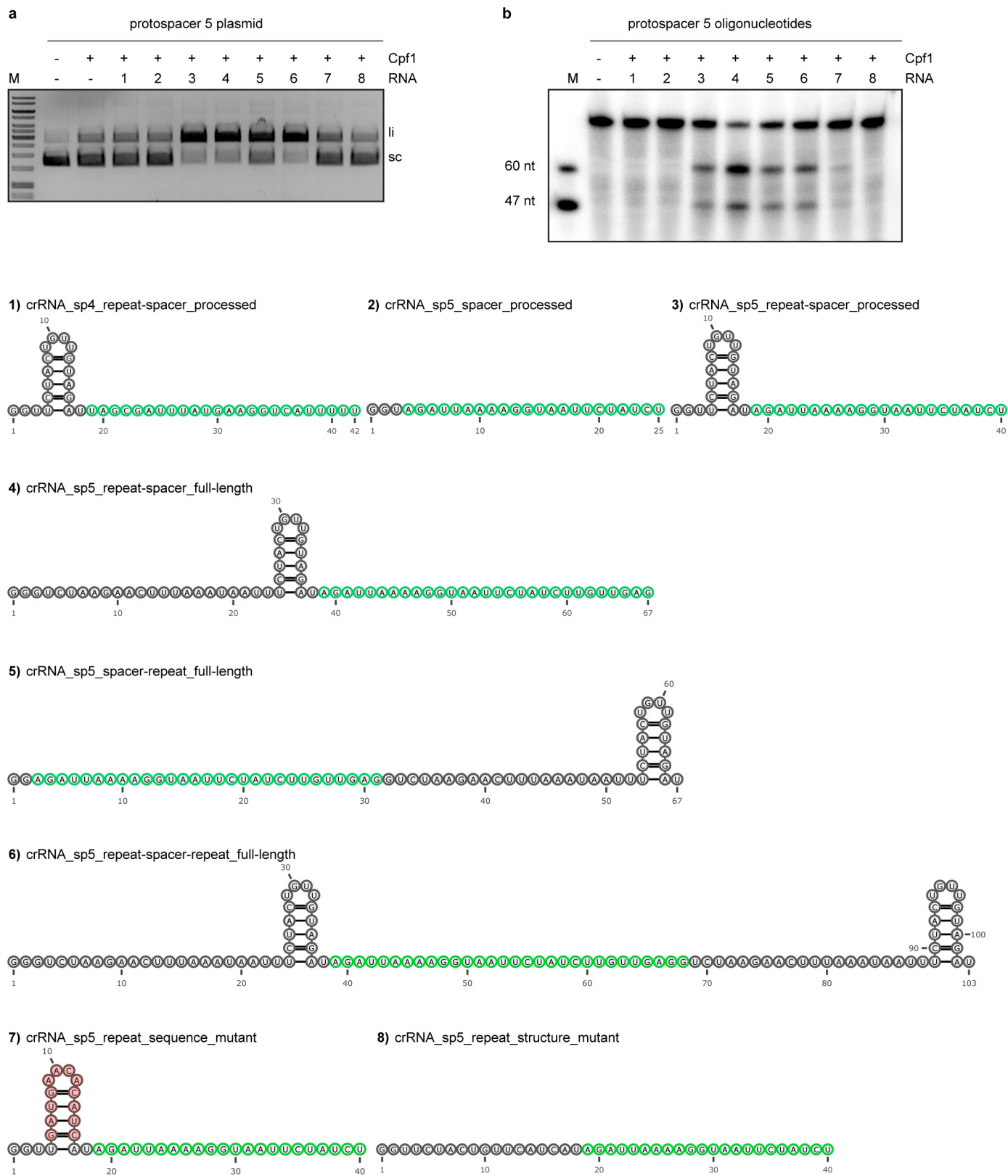
Extended Data Figure 4 | Cpf1 is a sequence- and structure-specific endoribonuclease. **a**, Design of various repeat variants of pre-crRNA-sp5 (pre-crRNA with spacer 5) with an altered repeat sequence, a destroyed repeat structure, single nucleotide exchanges (1–4) in the RRS (purple) and changed loop (green) and stem sizes (yellow). Note that the 5' repeat region of the wild-type repeat is not shown in the different variants. Red circles highlight the mutated or added residues. The RNA structures were generated with RNAfold⁴³ and visualized using VARNA⁴⁴ software. **b**, Internally labelled pre-crRNAs containing a wild-type repeat sequence, an altered repeat sequence or a destroyed repeat structure were obtained by *in vitro* transcription. The 5' end-labelled wild-type substrate was used to generate an alkaline hydrolysis ladder (OH) and an RNase T1 digest (T1) for size determination of the RNA fragments (Life Technologies). Cpf1 cleaved only the pre-crRNA template containing the wild-type repeat sequence yielding a small 19-nt 5' repeat fragment and a 50-nt

intermediate crRNA. **c**, Substrates with serial single mutations of the four RRS nucleotides (1–4, counting from the cleavage site) were tested for processing by Cpf1. Changes of the first three nucleotides were not tolerated for Cpf1-mediated processing, whereas changing the fourth nucleotide yielded a substrate that was processed with less efficiency compared to the wild-type substrate. **d**, The influence of loop variations in the repeat was tested with substrates containing +1 or –1 nucleotide in the loop. Both substrates were processed by Cpf1. Stems with +1 or –1 base pair, or +4 base pairs were used to determine length requirements of the stem. Cpf1 did not cleave any of the three substrates tested. The RNA cleavage reactions were performed by incubating 1 μ M of Cpf1 with 200 nM of RNA variant at 37 °C for 5 min in the presence of 10 mM MgCl₂. The cleavage products were analysed by denaturing polyacrylamide gel electrophoresis and phosphorimaging. RNA fragments are represented schematically and fragment sizes are indicated in nucleotides.



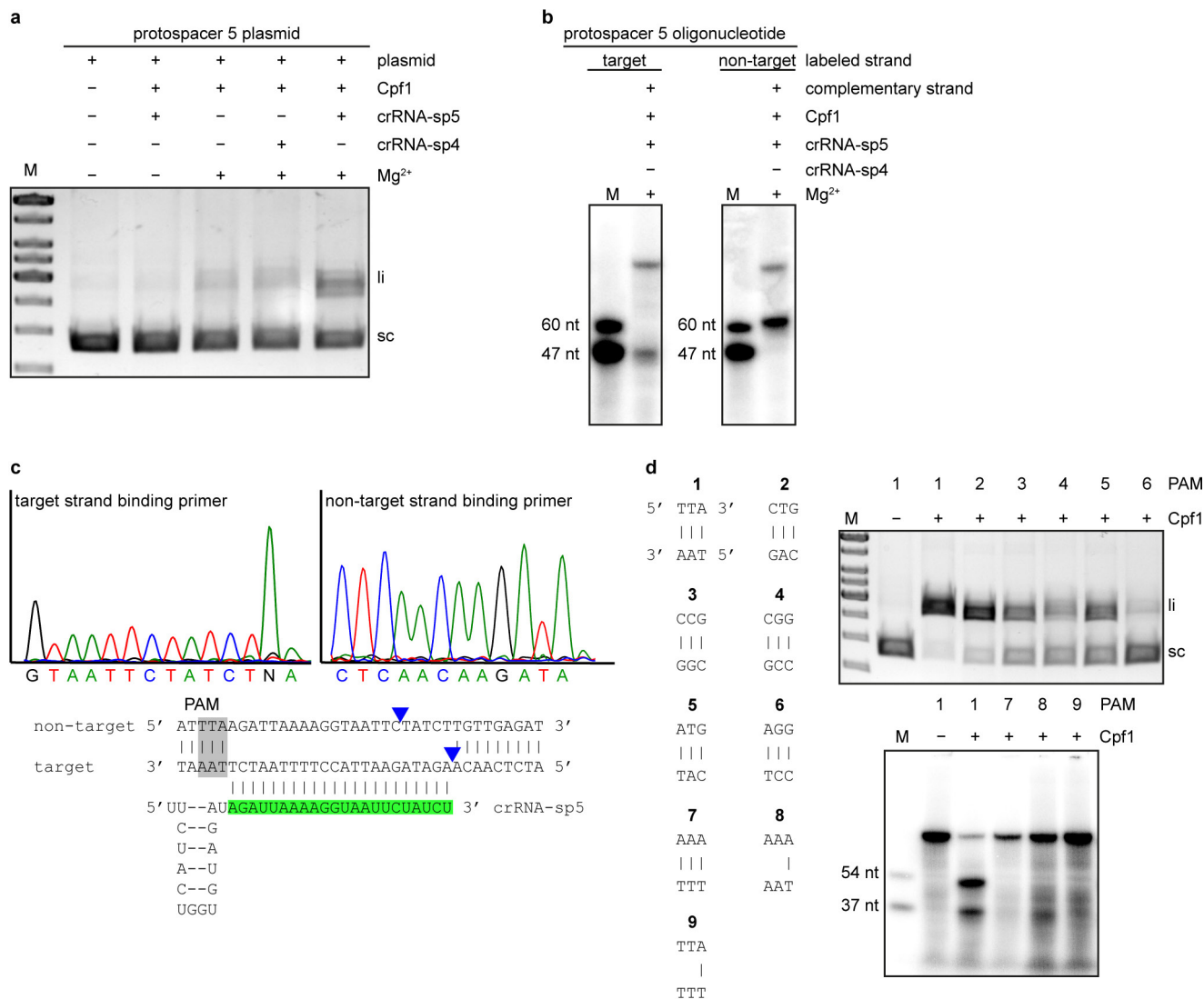
Extended Data Figure 5 | RNA and DNA cleavage activities of Cpf1 are dependent on divalent metal ions. **a**, Cleavage assays of pre-crRNA-sp5 (repeat-spacer 5, full-length, RNA 4, Extended Data Fig. 6) by Cpf1 in KGB buffer supplemented with different concentrations of divalent metal ion (indicated in mM) or EDTA (10 mM). Cleavage products were analysed by denaturing polyacrylamide gel electrophoresis and visualized by phosphorimaging. RNA fragments are represented schematically and fragment sizes are indicated in nucleotides. Specific RNA cleavage was observed in the presence of MgCl₂. Less specific cleavage was detected with CaCl₂, MnCl₂ and CoCl₂. No cleavage of pre-crRNA-sp5 was detected in presence of NiCl₂ and ZnCl₂. **b**, Cleavage assays of supercoiled

plasmid DNA containing protospacer 5 by Cpf1 programmed with crRNA-sp5 (repeat-spacer 5, processed, RNA3, Extended Data Fig. 6) in KGB buffer supplemented with different concentrations of divalent metal ions (indicated in mM). Cleavage products were analysed by agarose gel electrophoresis and visualized by EtBr staining. DNA cleavage was observed in the presence of MgCl₂ and MnCl₂. A more specific cleavage was observed in the presence of CaCl₂. The addition of CoCl₂, NiCl₂ or ZnCl₂ to the reaction did not result in DNA cleavage. li, linear; sc, supercoiled; M, 1 kb ladder (Fermentas). Quantification of three independent experiments shown in Extended Data Table 1b.



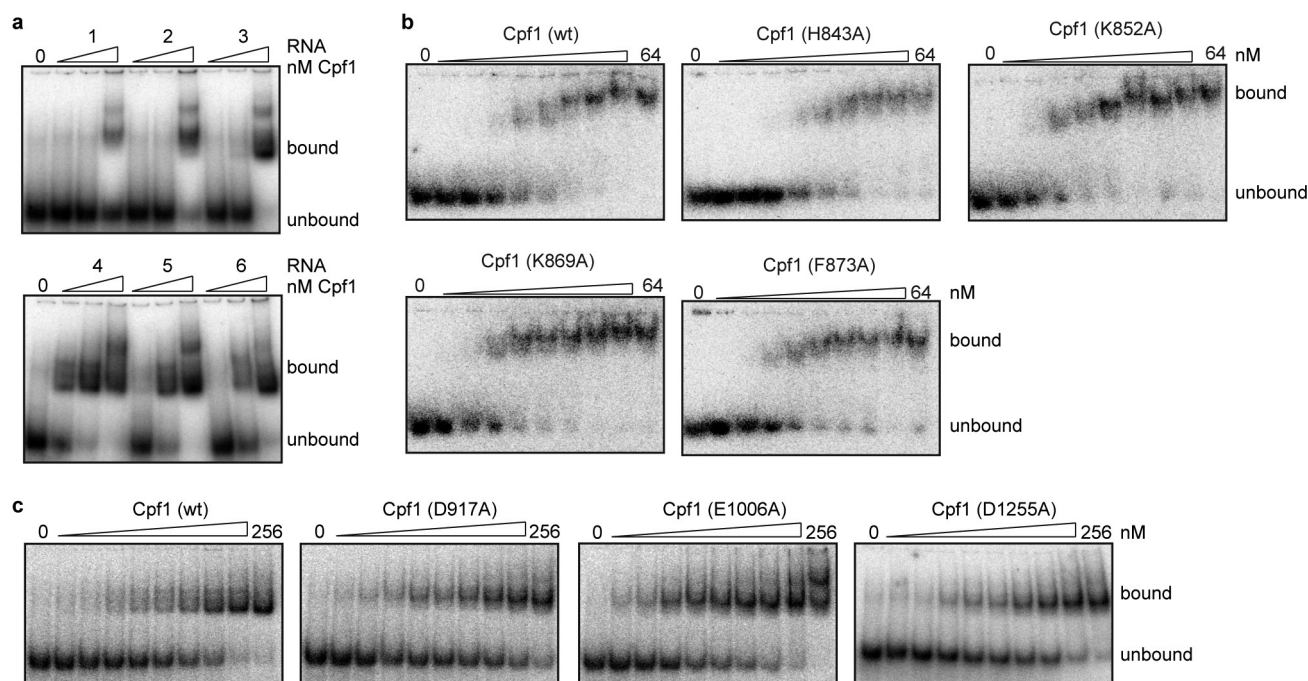
Extended Data Figure 6 | Cpf1 requires crRNA with an intact repeat structure to specifically cleave DNA. **a**, Cleavage assays of protospacer 5 containing supercoiled plasmid DNA by Cpf1 programmed with different RNA constructs (1, RNA construct containing spacer 4; 2–8, RNA constructs containing spacer 5) in the presence of 10 mM CaCl₂. Cleavage products were analysed by agarose gel electrophoresis and visualized by EtBr staining. **b**, Cleavage of 5'-radiolabelled oligonucleotide duplexes

containing protospacer 5 in the presence of 10 mM CaCl₂. Cleavage products were analysed by denaturing polyacrylamide gel electrophoresis and visualized by phosphorimaging. Fragment sizes are indicated in nucleotides. RNA structures were generated with RNAfold⁴³ and visualized using VARNA⁴⁴ software. Only the RNAs containing a full-length repeat and a spacer complementary to the target mediate DNA cleavage by Cpf1.



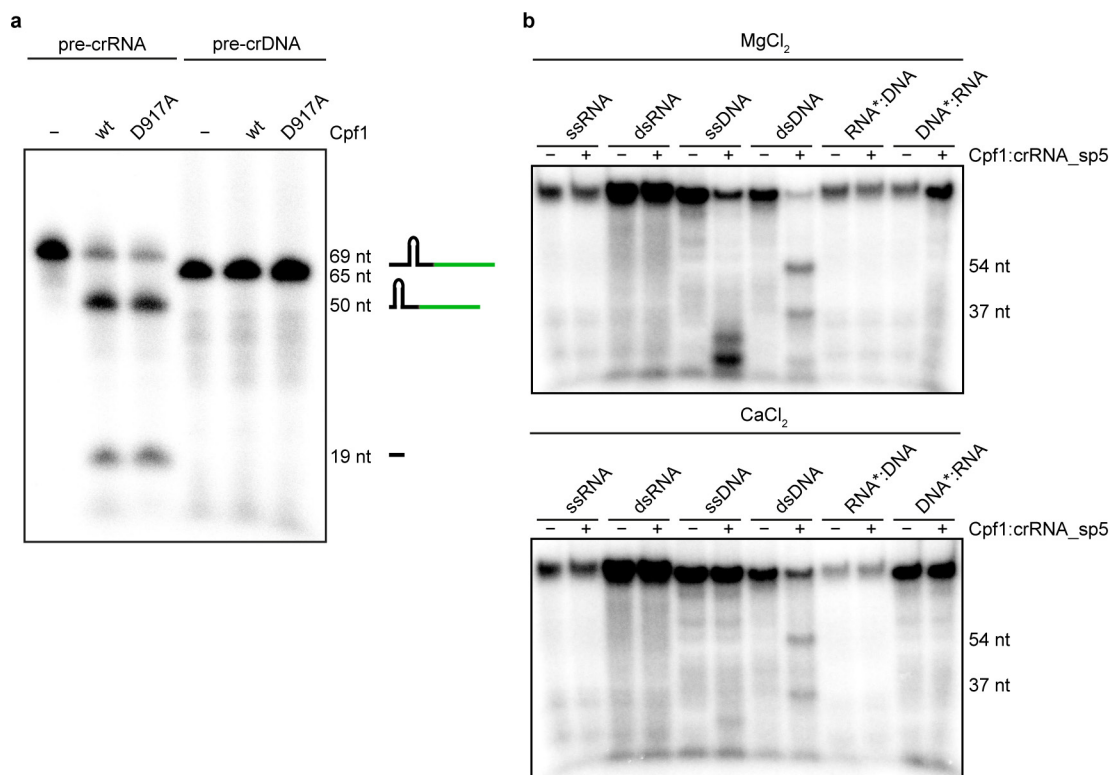
Extended Data Figure 7 | Analysis of target DNA cleavage by crRNA-programmed Cpf1 in presence of Mg²⁺. **a**, Cleavage assays of protospacer 5 containing supercoiled plasmid DNA by Cpf1 programmed with crRNA-sp4 or crRNA-sp5 (crRNA-sp4, repeat-spacer 4, processed, RNA 1, Extended Data Fig. 6; crRNA-sp5, repeat-spacer 5, processed, RNA 3, Extended Data Fig. 6) in absence or presence of 10 mM MgCl₂. Plasmid DNA cleavage was observed only with Cpf1 programmed with crRNA-sp5 in presence of Mg²⁺. **b**, Oligonucleotide cleavage assays using Cpf1 programmed with crRNA-sp5 (repeat-spacer 5, processed, RNA 3, Extended Data Fig. 6) in presence of 10 mM MgCl₂. Either the target or the non-target strand was 5'-radiolabelled before annealing to the

non-labelled complementary strand to form the duplex substrate. **c**, Sequencing analysis of the cleavage product obtained in **a**. The termination of the sequencing reaction indicates the cleavage site. Note that an enhanced signal for adenine is a sequencing artefact. **d**, Plasmid DNA containing protospacer 5 and the PAMs 1–6, or 5'-radiolabelled double-stranded oligonucleotide containing protospacer 5 and PAMs 1, 7–9 were subjected to cleavage by Cpf1 programmed with crRNA-sp5 (repeat-spacer, full-length, RNA 4, Extended Data Fig. 6) in the presence of 10 mM MgCl₂ (upper and lower panel, respectively). Oligonucleotide cleavage products are indicated in nucleotides.



Extended Data Figure 8 | Binding studies of Cpf1. **a**, EMSAs of 5'-radiolabelled double-stranded oligonucleotides containing protospacer 5 by Cpf1 programmed with RNA 1–6 (see Extended Data Fig. 6). The protein concentrations used were 8, 52 and 512 nM. Reactions were analysed by native PAGE and phosphorimaging. Unbound and bound DNAs are indicated. Higher DNA binding affinities are observed when Cpf1 is programmed with an RNA containing an entire repeat sequence. **b**, EMSAs of 5'-radiolabelled crRNA-sp5 (repeat-spacer 5, processed, RNA 3, Extended Data Fig. 6) by wild-type Cpf1, Cpf1(H843A), Cpf1(K852A), Cpf1(K869A) and Cpf1(F873A). The protein concentrations used were 2, 4, 8, 12, 16, 24, 32, 48 and 64 nM. Reactions were analysed by native polyacrylamide gel electrophoresis and phosphorimaging. Unbound and bound RNAs are indicated. Shown are representatives of at least three individual experiments. The bound and unbound RNA fractions were quantified, plotted against the enzyme concentration and fitted by nonlinear regression analysis. The calculated K_d values (\pm s.d.) were 16 ± 1 nM (wild type), 17 ± 0.5 nM (H843A), 12 ± 1 nM (K852A),

10 ± 1 nM (K869A) and 17 ± 1 nM (F873A). There are no differences between the RNA binding affinities of wild-type and mutant Cpf1. **c**, EMSAs of 5'-radiolabelled double-stranded oligonucleotides containing protospacer 5 targeted by wild-type Cpf1, Cpf1(D917A), Cpf1(E1006A) and Cpf1(D1255A) in complex with crRNA-sp5 (repeat-spacer 5, full-length, RNA 4, Extended Data Fig. 6). The protein concentrations used were 8, 16, 32, 42, 52, 64, 74, 128 and 256 nM. Reactions were analysed by native polyacrylamide gel electrophoresis and phosphorimaging. Unbound and bound DNAs are indicated. Shown are representative of at least three individual experiments. The bound and unbound DNA fractions were quantified, plotted against the enzyme concentration and fitted by nonlinear regression analysis. The calculated K_d values (\pm s.d.) were 50 ± 3 nM (wild type), 48 ± 8 nM (D917A), 40 ± 8 nM (E1006A) and 52 ± 6 nM (D1255A). There are no differences between the RNA-mediated DNA binding affinities of wild-type and mutant Cpf1. The reduced K_d for E1006A can be explained by the removal of the large negatively charged amino acid, which might facilitate interaction of Cpf1 with the DNA.



Extended Data Figure 9 | Processing activity of Cpf1 is specific for pre-crRNA and crRNA-mediated targeting of Cpf1 is directed only against single- and double-stranded DNA. **a**, Cpf1 processing activity was tested against pre-crRNA and pre-crDNA. Wild-type Cpf1 or Cpf1(D917A) (1 μ M) was incubated with 200 nM internally labelled pre-crRNA-sp5 (repeat-spacer 5, full-length, RNA 4, Extended Data Fig. 6) or a 5'-labelled ssDNA (pre-crDNA-sp5) construct with the same sequence as the RNA in KGB buffer with 10 mM MgCl₂ for 5 min at 37 °C. Incubation of wild-type Cpf1 and DNase inactive mutant (Cpf1(D917A)) with the RNA construct, but not the DNA construct, resulted in the expected cleavage products of a 19-nt repeat fragment and a 50-nt intermediate crRNA, indicating that the processing activity of Cpf1 is specific for RNA. **b**, crRNA-mediated DNA cleavage activity of Cpf1. Cpf1 (100 nM) in complex with crRNA-sp5

(repeat-spacer 5, full-length, RNA 4, Extended Data Fig. 6) was incubated with 10 nM of 5'-radiolabelled ssRNA, dsRNA, ssDNA, dsDNA or RNA-DNA hybrids in KGB buffer with either MgCl₂ (10 mM; upper panel) or CaCl₂ (10 mM; lower panel) for 1 h at 37 °C. The oligonucleotide DNA substrates contained the sequence for protospacer 5 targeted by the tested crRNA. For DNA-RNA hybrids, the 5'-radiolabelled target strand is indicated with an asterisk. Only ssDNA and dsDNA substrates were cleaved, indicating that the crRNA-mediated cleavage activity of Cpf1 is only directed against DNA substrates. The cleavage products for ssDNA, however, vary from those expected or observed for dsDNA. Cleavage reactions were analysed by denaturing polyacrylamide gel electrophoresis and phosphorimaging. RNA cleavage products are indicated schematically. RNA and DNA fragment sizes are given in nucleotides.

Extended Data Table 1

a. Quantification of Figure 2e.

substrate	wt	T22G	C21A	T20G	A19C	A18C	T17G	T16G	T15G
% cleavage	83 ± 15	37 ± 1	41 ± 2	22 ± 3	30 ± 2	33 ± 4	28 ± 11	39 ± 18	57 ± 2
substrate	T14G	C13A	C12A	A11C	T10G	T9G	A8C	A7C	G6T
% cleavage	69 ± 9	77 ± 13	87 ± 6	68 ± 12	79 ± 5	100 ± 0	65 ± 25	79 ± 16	92 ± 14
substrate	A5C	T4G	A3C	G2T	A1C	Mut_1-4		Mut_19-22	
% cleavage	75 ± 35	55 ± 27	62 ± 19	66 ± 24	64 ± 24	47 ± 25		0	

Percent cleavage is the result of three independent experiments ± standard deviation.

b. Quantification of Extended Data Figure 5b.

ion	Ca ²⁺		Mg ²⁺		Mn ²⁺		Co ²⁺		Ni ²⁺		Zn ²⁺	
concentration	1 mM	10 mM	1 mM	10 mM	1 mM	10 mM	1 mM	10 mM	1 mM	10 mM	1 mM	10 mM
% cleavage	44 ± 17	82 ± 8	13 ± 10	84 ± 10	39 ± 17	86 ± 2	0	0	0	0	0	0

Percent cleavage is the result of three independent experiments ± standard deviation.

The crystal structure of Cpf1 in complex with CRISPR RNA

De Dong^{1*}, Kuan Ren^{1*}, Xiaolin Qiu^{1*}, Jianlin Zheng¹, Minghui Guo¹, Xiaoyu Guan¹, Hongnan Liu¹, Ningning Li², Bailing Zhang¹, Daijun Yang¹, Chuang Ma¹, Shuo Wang¹, Dan Wu¹, Yunfeng Ma¹, Shilong Fan², Jiawei Wang², Ning Gao² & Zhiwei Huang¹

The CRISPR–Cas systems, as exemplified by CRISPR–Cas9, are RNA-guided adaptive immune systems used by bacteria and archaea to defend against viral infection^{1–7}. The CRISPR–Cpf1 system, a new class 2 CRISPR–Cas system, mediates robust DNA interference in human cells^{1,8–10}. Although functionally conserved, Cpf1 and Cas9 differ in many aspects including their guide RNAs and substrate specificity. Here we report the 2.38 Å crystal structure of the CRISPR RNA (crRNA)-bound *Lachnospiraceae* bacterium ND2006 Cpf1 (LbCpf1). LbCpf1 has a triangle-shaped architecture with a large positively charged channel at the centre. Recognized by the oligonucleotide-binding domain of LbCpf1, the crRNA adopts a highly distorted conformation stabilized by extensive intramolecular interactions and the (Mg(H₂O)₆)²⁺ ion. The oligonucleotide-binding domain also harbours a looped-out helical domain that is important for LbCpf1 substrate binding. Binding of crRNA or crRNA lacking the guide sequence induces marked conformational changes but no oligomerization of LbCpf1. Our study reveals the crRNA recognition mechanism and provides insight into crRNA-guided substrate binding of LbCpf1, establishing a framework for engineering LbCpf1 to improve its efficiency and specificity for genome editing.

After integration of short segments of invader-derived DNA (known as a protospacer) into a CRISPR array within the host genome, expression and processing of the precursor crRNAs produces mature crRNAs. The mature crRNAs then guide an effector protein, either a large single Cas protein (class 2 CRISPR systems) or a Cas protein complex (class 1 CRISPR systems), to target and cleave foreign DNAs (or RNAs in some cases) bearing complementary sequences^{7,10–13}. Typical examples of class 2 CRISPR systems include the well-characterized CRISPR–Cas9^{1,10}. The combination of Cas9 from *Streptococcus pyogenes* (SpyCas9) and a synthetic single-guide RNA (sgRNA) that contains a guide region and duplex of crRNA and *trans*-activating crRNA (tracrRNA) has been harnessed as a two-component programmable system for genetic manipulation of various organisms^{14–16}.

Cas9 as an endonuclease is a modular protein, comprising an RNA-recognizing domain and two nuclease domains (HNH and RuvC) connected by an arginine-rich bridge helix, and a protospacer-adjacent motif (PAM, a short sequence located immediately downstream of the target DNA sequence)-interacting domain^{17–19}. Recognition of the sgRNA induces marked conformational rearrangement of Cas9 (refs 20, 21), recruiting target double-stranded (ds)DNA through pairing of the guide region from the sgRNA with the dsDNA^{17–21}. In the sgRNA–Cas9–target DNA ternary complex, the two strands of dsDNA complementary and non-complementary to the guide segment of sgRNA are cleaved by the HNH and RuvC nuclease domains, respectively^{22,23}.

In the CRISPR–Cpf1 system, a single 42–44 nucleotide (nt) crRNA with ~19 nt direct repeat sequence followed by 23–25 nt of spacer

sequence is sufficient to guide the endonuclease Cpf1 (~1,300 residues) for dsDNA targeting⁹. This is distinctly different from the CRISPR–Cas9 system wherein much longer guide RNAs, such as the sgRNA with ~80–100 nt, as well as the crRNA–tracrRNA duplex and the secondary structure of the tracrRNA 3' end are required for DNA recognition and cleavage^{19,21,22,24}. Only a single nuclease domain (RuvC), required for its dsDNA cleavage activity⁹, is identifiable in Cpf1. Therefore, whether Cpf1 acts as a dimer to cleave the two strands of dsDNA, or another unknown active site tightly coupled to that of RuvC present in Cpf1, remains unknown. Interestingly, unlike Cas9 that generates cleavage products with blunt ends^{23,25}, Cpf1 makes staggered cuts leaving five-nucleotide 5' overhangs distal to the PAM site⁹. Moreover, Cpf1 utilizes a T-rich PAM sequence, in contrast to the G-rich PAM preference of Cas9 (ref. 16).

To understand how LbCpf1 recognizes crRNA, we solved the crystal structure of full-length LbCpf1 in complex with a 43-nt crRNA at 2.38 Å resolution by the single-wavelength anomalous dispersion method using a SeMet-derived protein (Fig. 1a, b, and Extended Data Table 1). The overall structure of the LbCpf1–crRNA binary complex is bilobal, but assumes a triangle-shaped architecture with a large positively charged channel at the centre (Fig. 1c). The structure of the crRNA-bound LbCpf1 can be divided into three portions: the N-terminal helical domain, the central oligonucleotide-binding domain (OBD) and the C-terminal RuvC domain (Fig. 1a and b). The OBD and RuvC juxtapose with each other, forming one side of the triangle architecture (Fig. 1b). The N-terminal helical domain can be further divided into helical I (H1) and helical II (H2) that pack loosely against each other, forming another side of the triangle-shaped structure (Fig. 1b and d). The RuvC domain of LbCpf1 (Cpf1^{RuvC}) comprises three RuvC motifs (RuvC-I–III). The structures of LbCpf1^{RuvC}, *Staphylococcus aureus* (Sa) Cas9^{RuvC} (ref. 18) and SpyCas9^{RuvC} (ref. 17) can be well aligned, with r.m.s.d. of 3.7 Å for 161 equivalent Cα atoms between LbCpf1^{RuvC} and SaCas9^{RuvC} (PDB, 5CZZ), and r.m.s.d. of 4 Å for 150 equivalent Cα atoms between LbCpf1^{RuvC} and SpyCas9^{RuvC} (PDB, 4UN3) (Extended Data Fig. 1). Importantly, the catalytic residues of the three RuvC domains are well superimposed (Extended Data Fig. 1). An embedded domain within the LbCpf1^{RuvC} is formed through packing of a four helical bundles against three antiparallel strands (Fig. 1b). Searching in the Dali database did not identify any structures appreciably homologous with this domain, indicating that it represents a novel fold (domain with unknown functions, termed 'UK' domain). Interaction of this LbCpf1 domain with H2 results in the formation of the third side of the triangle (Fig. 1b and d). One β-strand formed by the extreme N-terminal side of H1 pairs with one strand from the OBD, forming one vertex of the structure (Fig. 1b and e). A looped-out helical domain (LHD) from the OBD is positioned nearly perpendicular to the planar triangle (Fig. 1b and d).

¹School of Life Science and Technology, Harbin Institute of Technology, Harbin 150080, China. ²Ministry of Education Key Laboratory of Protein Sciences, Center for Structural Biology, School of Life Sciences, Tsinghua University, Beijing 100084, China.

*These authors contributed equally to this work.

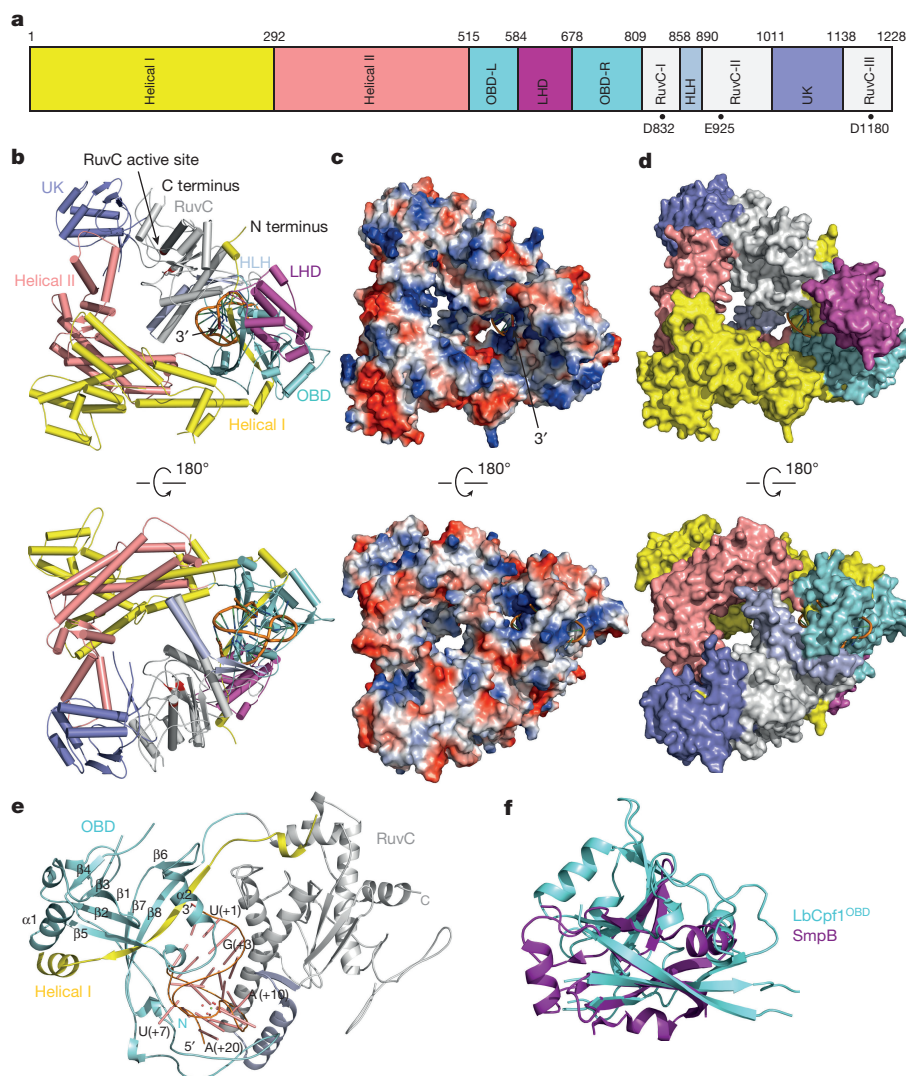


Figure 1 | Triangle-shaped structure of the LbCpf1-crRNA complex.

a, Schematic diagram of domain organization of LbCpf1. **b**, Overall structures of the LbCpf1-crRNA complex shown in two different orientations. The position of the active site of RuvC domain is indicated. Individual LbCpf1 domains are coloured according to the scheme in **a**. crRNA is shown in cartoon and coloured orange. **c**, **d**, Electrostatic

potential surface (**c**) and surface representations (**d**) of LbCpf1, shown in the same orientations as in **a**. **e**, crRNA interacts with OBD and RuvC of LbCpf1. The secondary structural elements of the OBD are labelled. **f**, Structural superposition of LbCpf1^{OBD} (aquamarine) with small protein B (PDB, 1WJX; magenta)²⁹.

The LbCpf1^{OBD} consists of a central β -barrel (including the β -strand from LbCpf1^{H1}) enclosed by three helices on one side and four on the opposite other side (Fig. 1e). Searching in the Dali database showed that LbCpf1^{OBD} is structurally homologous with the RNA-binding protein, small protein B (PDB, 1WJX), with an r.m.s.d. of 3.58 Å over 77 aligned C α atoms (Fig. 1f), that has a typical oligonucleotide-binding fold conserved in other RNA-binding proteins associated with the ribosome²⁶. The crRNA is mainly sandwiched between a long hairpin loop and two helices of LbCpf1^{OBD} (Fig. 1e). Further strengthening LbCpf1-crRNA interaction, one β -hairpin loop and the following α -helix connecting RuvC-I and RuvC-II motifs of LbCpf1^{RuvC} also contact another surface of the crRNA (Fig. 1e). The 3' end of the LbCpf1-bound crRNA is poised to point into the central channel between the LbCpf1^{RuvC} and LbCpf1^{H1} (Fig. 1b and c).

The direct sequence of the LbCpf1-bound crRNA (Fig. 2a), well defined in the electron density map, adopts a highly distorted fold containing a short stem-loop-like structure (Fig. 2b). The crRNA conformation is markedly different from that of the sgRNA bound by SpyCas9 or SaCas9 (Extended Data Fig. 2) and stabilized through extensive intramolecular interactions. The stem-loop formed by five Watson-Crick base pairs appears to be important for crRNA to keep its

conformation in LbCpf1 (Fig. 2c). The crRNA stem is further strengthened by a base pair made between U(+1) and U(+17) (Fig. 2c). The nucleobase of U(+18) inserts deeply into the centre of the stem-loop and forms hydrogen bonds with the stem-loops of A(+13), U(+14) and C(+15) (Fig. 2d) to stabilize the conformation of the crRNA. Supporting a role for these intramolecular hydrogen bonds in crRNA recognition by LbCpf1, mutations of U(+18) resulted in loss of the dsDNA cleavage activity of Cpf1⁹. A(+19) pairs with U(+11) and stacks against G(+6) (Fig. 2d), further stabilizing the smaller U-shaped structure.

Electron density with a clearly octahedral shape is located at the centre of the crRNA (Fig. 2b and e). As Mg²⁺ is the only metal ion contained in the crystallization buffer, this structural observation is reminiscent of the ribozymes and other RNA/DNA duplexes that contain an octahedral (Mg(H₂O)₆)²⁺ ion for stabilization of their specific conformations²⁷. Building the hydrated Mg²⁺ into the density allowed A(+13), U(+14), U(+18), A(+19) and A(+20) to coordinate with the ion (Fig. 2e). The crRNA conformation stabilized by intramolecular interactions and (Mg(H₂O)₆)²⁺ may be important for its recognition by LbCpf1. Thus, the Mg²⁺-dependent endonuclease activity of *Francisella* FnCpf1 (ref. 9) could partially result from stabilization of the crRNA

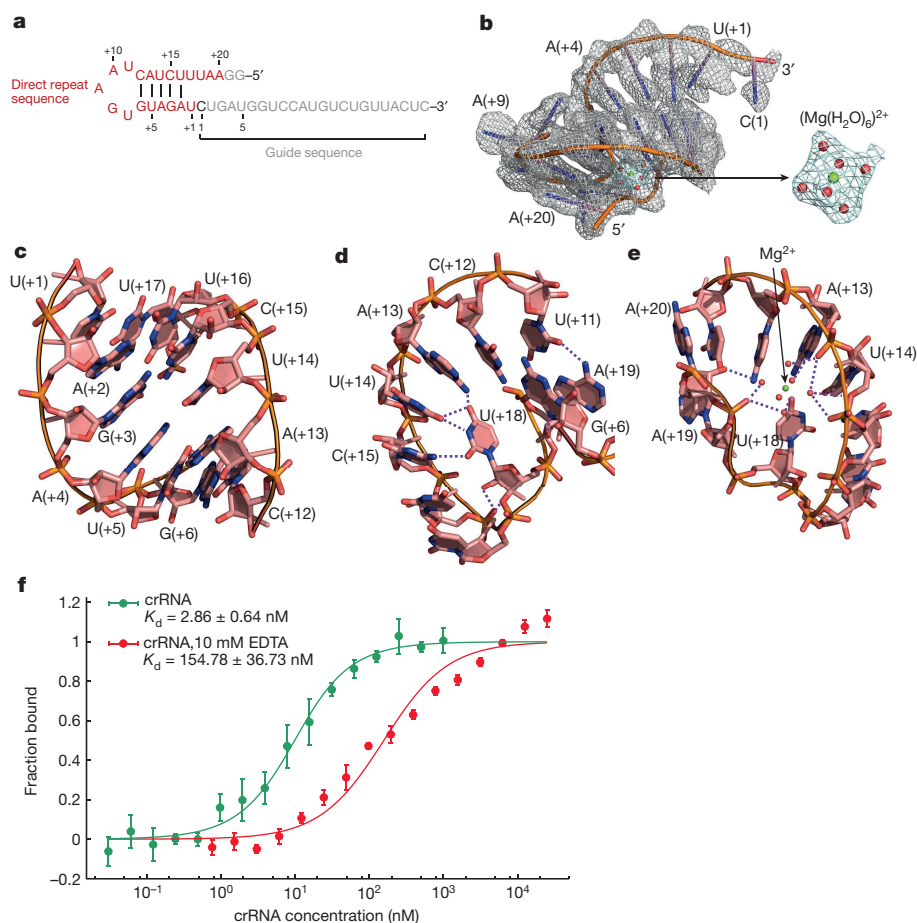


Figure 2 | The distorted conformation of the LbCpf1-bound crRNA is stabilized by extensive intramolecular interactions and $(\text{Mg}(\text{H}_2\text{O})_6)^{2+}$.

a, Schematic of the crRNA used for structural analysis. **b**, Cartoon representation of the structure of the LbCpf1-bound crRNA. Shown in mesh (grey) is the electron density $2F_o - F_c$ map surrounding the crRNA, contoured at 1.2σ . The nucleotides of crRNA are labelled. **c**, Close-up views of the stem region of crRNA. **d**, Detailed interactions of the 5' side

of crRNA with the stem loop. Hydrogen bonds are shown as dashed lines. **e**, Interactions between the $(\text{Mg}(\text{H}_2\text{O})_6)^{2+}$ ion and crRNA. Mg^{2+} and its coordinated water molecules are indicated by green and red spheres, respectively. **f**, EDTA-treated crRNA displays a comprised LbCpf1-binding activity. Data shown are representative of three independent microscale thermophoresis experiments in the absence (green) and presence (red) of EDTA. Error bars, s.d. ($n = 3$).

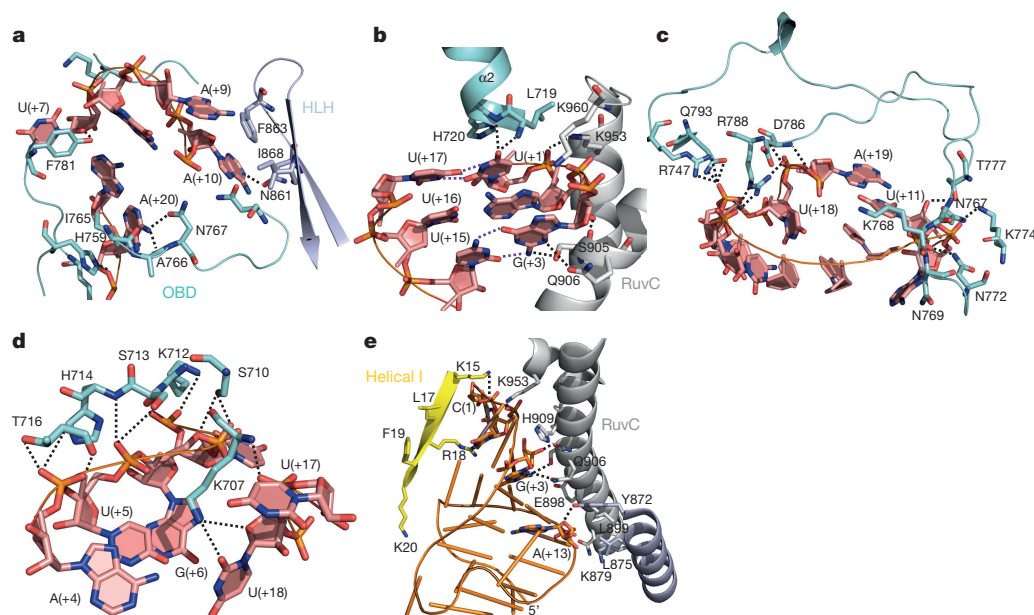


Figure 3 | Recognition of crRNA by LbCpf1. **a**, **b**, Interactions of crRNA nucleobases with LbCpf1. The side chains from the OBD (aquamarine), HLH (lightblue) and RuvC (grey) are labelled. Purple and black dashed lines represent intra- and intermolecular hydrogen bonds, respectively.

c, **d**, Detailed interactions of sugar-phosphate backbone of crRNA with LbCpf1 OBD. The side chains from the OBD are labelled and shown in aquamarine. **e**, Interactions of crRNA with LbCpf1^{RuvC} (grey) and LbCpf1^{HI} (yellow).

conformation by the metal ion. If this is the case, removal of Mg^{2+} is expected to compromise crRNA binding to LbCpf1. Indeed, the binding affinity between crRNA and LbCpf1 was reduced ~ 50 -fold as compared to that in the absence of EDTA (Fig. 2f). Other hydrated metal ions with a similar geometry to the $(Mg(H_2O)_6)^{2+}$ ion may also function to stabilize the conformation of crRNA.

Only the direct repeat sequence was well defined in the final refined electron density (Fig. 2a and b). This was not caused by degradation of the crRNA, because a TBE-urea PAGE analysis showed that the crRNA in the crystals remained intact (Extended Data Fig. 3a). Interestingly, removal of the guide sequence slightly compromised the binding affinity between the crRNA and LbCpf1, although it is not involved in their interaction (Extended Data Fig. 3b). Further supporting these results, the guide-sequence-truncated crRNA (crRNA*) inhibited crRNA-induced dsDNA cleavage by LbCpf1 (Extended Data Fig. 3c) in a dose-dependent manner.

The distorted crRNA makes extensive interactions with LbCpf1 and most of them are mediated by the sugar–phosphate backbone of the bound crRNA (Fig. 1e). But the nucleobases of A(+20), A(+10), U(+7), G(+3) and U(+1), most of which are splayed out, also contribute to crRNA recognition by LbCpf1 (Fig. 3a and b). As well as stabilization of the conformation of crRNA by forming non-conventional pairing with U(+17) and stacking against LbCpf1^{His720} (Fig. 3b), U(+1) forms a pair of hydrogen bonds with the amide nitrogen atoms of LbCpf1^{Leu719His720} at the N-terminal side of helix $\alpha 2$ from OBD. These structural observations explain the requirement for uracil at this position for the activity of crRNA⁹. The three U(+1)-interacting residues are well conserved among Cpf1 orthologues (Extended Data Fig. 4), suggesting a similar role for them in crRNA recognition. In contrast, A(+9), a position highly variable among 16 Cpf1 family crRNAs⁹ (Fig. 3a), is completely solvent-exposed and not involved in interactions with LbCpf1. This explains why these crRNAs are exchangeable as substrates of FnCpf1 (ref. 9).

The long hairpin loop connecting $\beta 7$ and $\beta 8$ wraps halfway around the crRNA. Either side or main chains of many residues from this loop form polar interactions with the sugar–phosphate backbone (Fig. 3c and d). As well as polar interactions, van der Waals contacts also contribute to H1 and RuvC interaction with the crRNA (Fig. 3e). Additionally, several 2'-OH groups are involved in interactions with LbCpf1 (Fig. 3b and c), explaining why LbCpf1 specifically recognizes RNA. Structure-based sequence alignment revealed that the crRNA-interacting residues are largely conserved among LbCpf1 protein from other species, suggesting that they have a conserved crRNA recognition mechanism (Extended Data Fig. 4).

No homodimer or higher order oligomer of the LbCpf1–crRNA complex was detected in the crystals. Consistently, analytical ultracentrifugation showed that the LbCpf1 protein exhibited a molecular weight of ~ 127 kDa in the absence of crRNA and ~ 145 kDa in the presence of crRNA (Extended Data Fig. 5a). These results indicate that LbCpf1 was monomeric in solution and crRNA binding induced no LbCpf1 oligomerization. Interestingly, addition of 45 nt of crRNA greatly reduced the frictional coefficient of LbCpf1 (Extended Data Fig. 5a), suggesting that crRNA binding resulted in a more compact conformation of the LbCpf1 protein. Indeed, crRNA rendered LbCpf1 much less sensitive to degradation by trypsin (Extended Data Fig. 5b). Similar results were obtained with the 22 nt direct repeat sequence of crRNA (crRNA*). This is markedly different from the SpyCas9-bound sgRNA that requires both the direct repeat and guide sequences to alter LbCpf1 conformation²¹. Addition of dsDNA complementary to crRNA did not further change the degradation pattern of LbCpf1 by trypsin (Extended Data Fig. 5b). Negative staining electron microscopy showed that particles of LbCpf1 protein in the presence or absence of crRNA were monomeric in solution (Extended Data Fig. 5c). However, in the absence of RNA, LbCpf1 displayed extended conformations and appeared more structurally dynamic as indicated by two-dimensional averages (Extended Data Fig. 5c). In contrast,

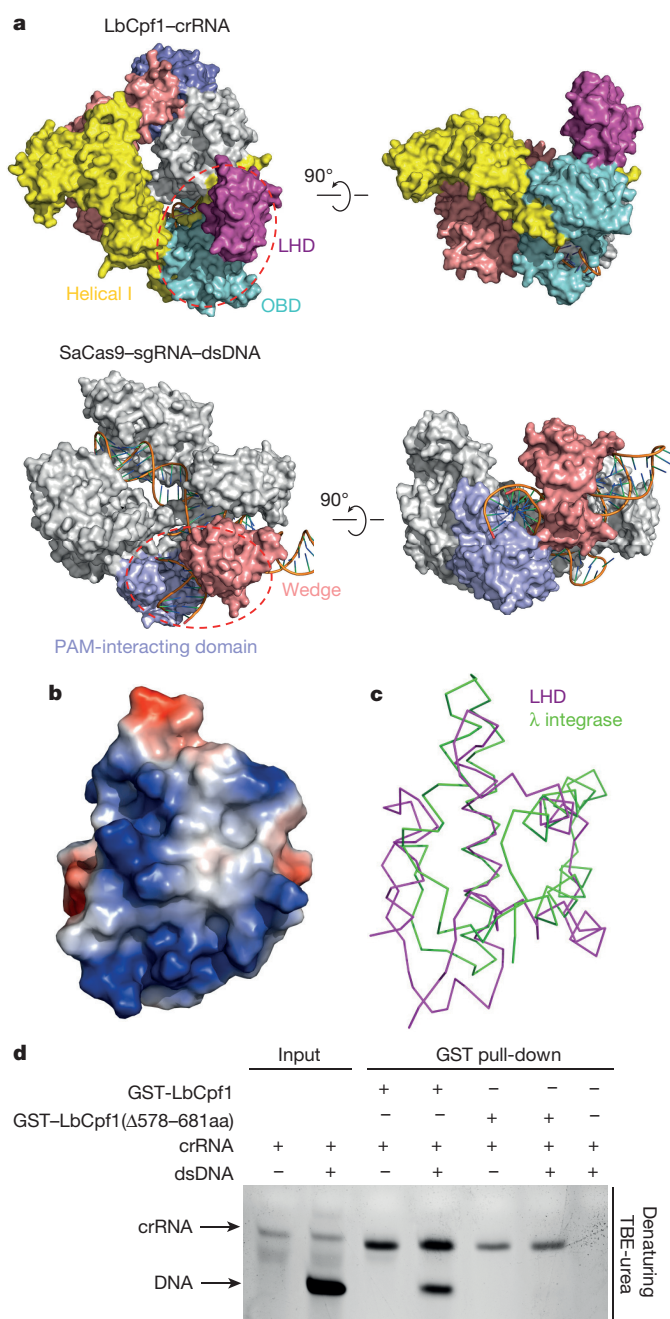


Figure 4 | LHD is involved in dsDNA binding of LbCpf1. **a**, Structural comparison of LbCpf1–crRNA and SaCas9–sgRNA–dsDNA (PDB, 5CZZ) shown in two different orientations. The putative dsDNA-binding region of LbCpf1 and the dsDNA-bound interface on SaCas9 are highlighted. **b**, Electrostatic potential surface representation of LbCpf1^{LHD}. White, blue and red indicate neutral, positive and negative surfaces respectively. **c**, LbCpf1^{LHD} displays structural homology with the bacteriophage λ integrase. Structural superimposition between LbCpf1^{LHD} and λ integrase (PDB, 2OXO; green)³⁰. **d**, LbCpf1 mutant lacking the LHD fails to bind dsDNA *in vitro*. A GST-fused inactive LbCpf1 (D832A/E925A) or LHD-truncated LbCpf1 proteins was first bound to glutathione-sepharose beads and incubated with crRNA or dsDNA as indicated. After extensive washing, the bound nucleotides were run on TBE-urea polyacrylamide gels and visualized by ethidium bromide staining. Data shown is representative of three independent experiments. Non-target strand/sequence: 5'-CTTTAGAGAAGTCATTTAATAAGGCCACTG-3'.

addition of crRNA markedly changed the elongated particles to more compact triangle-shaped structures (Extended Data Fig. 5c), as observed in the crystal structure. Collectively, our results show

that crRNA binding drives pronounced conformational changes in LbCpf1.

The PAM-interacting domain is necessary for the CRISPR–Cas9 systems to unwind dsDNA substrates and thereby form a crRNA–Cas9–dsDNA ternary complex^{17–19}. The PAM duplex of the dsDNA also interacts with a functionally uncharacterized domain called the wedge domain (Fig. 4a). Interestingly, the wedge domain of SaCas9 is located at an equivalent position to the looped-out helical domain of LbCpf1 (Fig. 4a). Additionally, the looped-out helical domain is highly positively charged on the side facing the central channel of LbCpf1 (Fig. 1c and 4b). Dali search revealed that this structural domain shares appreciable similarity with the bacteriophage λ integrase (PDB, 2OXO; 17% identity, r.m.s.d. of 2.9 Å for equivalent 42 C α atoms) (Fig. 4c). These data collectively suggest that the looped-out helical domain may be involved in interactions with dsDNA substrates of LbCpf1. This LbCpf1 domain, which is stabilized by interaction with its neighbouring LbCpf1 molecules (Extended Data Fig. 6), makes no contact with any other parts of LbCpf1 in the structure. Thus, its removal is predicted to have no effect on the structural integrity of the remaining part of LbCpf1. Indeed, an LbCpf1 mutant with this structural domain deleted completely lost dsDNA binding activity (Fig. 4d), but exhibited a strong affinity with crRNA (Fig. 4d and Extended Data Fig. 7).

Cpf1 also functions as an RNase to process pre-crRNAs for maturation^{9,28}. H843, K852 and K869 from FnCpf1 were found to be important for the RNA processing activity of FnCpf1 (ref. 28). The nitrogen atoms from the side chains of these three conserved residues (H759, K768 and K785 in LbCpf1) are co-planar and form hydrogen bonds with the phosphate group from the processed site A(+20) (Extended Data Fig. 8), suggesting that processing of RNA may be a base-catalysed reaction.

Our electron microscopy and biochemical studies showed that crRNA binding induced pronounced structural rearrangements of LbCpf1, leading to formation of a substrate-binding conformation of LbCpf1 (Extended Data Fig. 9a). Further conformational changes accompanied by binding of a LbCpf1 dsDNA substrate are still possible. Given the orientation of the 3' end of the LbCpf1-bound crRNA, it is reasonable to assume that the positively charged central channel is the site where the heteroduplex formed between crRNA and substrate DNA binds (Extended Data Fig. 9b). We provide evidence for the involvement of the looped-out helical domain in substrate binding. This structural domain together with the portion of H1 adjacent to the crRNA, positioned similarly to the wedge and the PAM-interacting domain in SaCas9 (Fig. 4a), could function as the DNA duplex binding site.

In summary, the data presented here reveal the crRNA recognition mechanism of LbCpf1 and provide insight into crRNA-induced substrate binding of this CRISPR protein. Our study opens the possibility of engineering Cpf1, which is expected to generate Cpf1 mutants with more efficiency and specificity for genome manipulation and even therapeutic applications.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 22 February; accepted 29 March 2016.

Published online 20 April 2016.

1. Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR–Cas systems. *Nature Rev. Microbiol.* **13**, 722–736 (2015).
2. Wiedenheft, B., Sternberg, S. H. & Doudna, J. A. RNA-guided genetic silencing systems in bacteria and archaea. *Nature* **482**, 331–338 (2012).
3. Marraffini, L. A. CRISPR–Cas immunity in prokaryotes. *Nature* **526**, 55–61 (2015).
4. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).

5. Garneau, J. E. *et al.* The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67–71 (2010).
6. Westra, E. R. *et al.* The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity. *Annu. Rev. Genet.* **46**, 311–339 (2012).
7. Sorek, R., Lawrence, C. M. & Wiedenheft, B. CRISPR-mediated adaptive immune systems in bacteria and archaea. *Annu. Rev. Biochem.* **82**, 237–266 (2013).
8. Schunder, E., Rydzewski, K., Grunow, R. & Heuner, K. First indication for a functional CRISPR/Cas system in *Francisella tularensis*. *Int. J. Med. Microbiol.* **303**, 51–60 (2013).
9. Zetsche, B. *et al.* Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR–Cas system. *Cell* **163**, 759–771 (2015).
10. Wright, A. V., Nuñez, J. K. & Doudna, J. A. Biology and applications of CRISPR systems: harnessing nature's toolbox for genome engineering. *Cell* **164**, 29–44 (2016).
11. Barrangou, R. & Marraffini, L. A. CRISPR–Cas systems: prokaryotes upgrade to adaptive immunity. *Mol. Cell* **54**, 234–244 (2014).
12. Horvath, P. & Barrangou, R. CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**, 167–170 (2010).
13. van der Oost, J. *et al.* CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem. Sci.* **34**, 401–407 (2009).
14. Jiang, W. & Marraffini, L. A. CRISPR–Cas: new tools for genetic manipulations from bacterial immunity systems. *Annu. Rev. Microbiol.* **69**, 209–228 (2015).
15. Sternberg, S. H. & Doudna, J. A. Expanding the biologist's toolkit with CRISPR–Cas9. *Mol. Cell* **58**, 568–574 (2015).
16. Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR–Cas9 for genome engineering. *Cell* **157**, 1262–1278 (2014).
17. Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513**, 569–573 (2014).
18. Nishimasu, H. *et al.* Crystal structure of *Staphylococcus aureus* Cas9. *Cell* **162**, 1113–1126 (2015).
19. Nishimasu, H. *et al.* Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935–949 (2014).
20. Jinek, M. *et al.* Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* **343**, 1247997 (2014).
21. Jiang, F. *et al.* A Cas9-guide RNA complex preorganized for target DNA recognition. *Science* **348**, 1477–1481 (2015).
22. Deltcheva, E. *et al.* CRISPR RNA maturation by *trans*-encoded small RNA and host factor RNase III. *Nature* **471**, 602–607 (2011).
23. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
24. Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnol.* **31**, 827–832 (2013).
25. Garneau, J. E. *et al.* The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67–71 (2010).
26. Draper, D. E. & Reynaldo, L. P. RNA binding strategies of ribosomal proteins. *Nucleic Acids Res.* **27**, 381–388 (1999).
27. Draper, D. E., Grilley, D. & Soto, A. M. Ions and RNA folding. *Annu. Rev. Biophys. Biomol. Struct.* **34**, 221–243 (2005).
28. Fonfara, I. *et al.* The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature* <http://dx.doi.org/10.1038/nature17945> (20 April 2016).
29. Bessho, Y. *et al.* Structural basis for functional mimicry of long-variable-arm tRNA by transfer-messenger RNA. *Proc. Natl Acad. Sci. USA* **104**, 8293–8298 (2007).
30. Kamadurai, H. B., Jain, R. & Foster, M. P. Crystallization and structure determination of the core-binding domain of bacteriophage lambda integrase. *Acta Crystallogr. Sect. F* **64**, 470–473 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank F. Yu and J. He at Shanghai Synchrotron Radiation Facility for help with data collection. We thank J. Chai for critical reading of the manuscript. We acknowledge the Tsinghua University Branch of China National Center for Protein Sciences Beijing for providing the facility support. This research was funded by the National Natural Science Foundation of China grant numbers 31422014, 31450001 and 31300605 to Z.H.

Author Contributions Z.H. designed the experiments. D.D., K.R. and X.Q. performed the bulk of the experiments. Data were analysed by Z.H., D.D., K.R. and X.Q.; J. Z., M.G., X.G., H. L., N.L., D.Y., C.M., S.W., D.W., B.Z., Y.M., S.F., N.G. and J.W. contributed to some experiments and discussions. Z.H., D.D., K.R. and X.Q. wrote the paper.

Author Information The atomic coordinates and structure factors of the LbCpf1–crRNA complex have been deposited in the Protein Data Bank under the accession code 5ID6. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Z.H. (huangzhiwei@hit.edu.cn).

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Protein expression and purification. The cDNA of full-length LbCpf1 was synthesized and sub-cloned into the expression vector pGEX-6P-1 (with an N-terminal GST tag and a precision protease cleavage site between GST and LbCpf1). The LbCpf1 protein was expressed in *E. coli* C43 (DE3) cells (BioVector NTCC). Expression of the recombinant protein was induced by 0.3 mM isopropyl β -D-1-thiogalactopyranoside (IPTG) at 16°C. After overnight induction, the cells were collected by centrifugation, resuspended in buffer A (25 mM Tris-HCl, pH 8.0, 1 M NaCl, 3 mM DTT, 1 mM MgCl₂) supplemented with 1 mM protease-inhibitor PMSF (phenylmethanesulphonylfluoride, Sigma). The cells were subjected to lysis by sonication and cell debris was removed by centrifugation at 23,708 g for 40 min at 4°C. The lysate was first purified using glutathione sepharose 4B (GS4B) beads (GE Healthcare). The beads were washed and the bound proteins were cleaved by precision protease in buffer B (25 mM Tris-HCl, pH 8.0, 150 mM NaCl, 3 mM DTT, 1 mM MgCl₂) overnight at 4°C to remove the GST tag. The eluted LbCpf1 protein was further fractionated by heparin sepharose column and ion-exchange chromatography and finally cleaned by size-exclusion chromatography (HiLoad 16/600 Superdex200, GE Healthcare) with buffer C (10 mM Tris-HCl, pH 8.0, 150 mM NaCl, 3 mM DTT) via FPLC (AKTA Pure). Selenomethionine (SeMet)-substituted LbCpf1 was expressed in *E. coli* C43 (DE3) cells grown in M9 minimal medium supplemented with 60 mg l⁻¹ SeMet (Sigma-Aldrich) and specific amino acids: Ile, Leu and Val at 50 mg l⁻¹; Lys, Phe and Thr at 100 mg l⁻¹. The SeMet LbCpf1 protein was purified as described above.

To assemble the LbCpf1–crRNA complex, LbCpf1 protein was incubated with crRNA at the molar ratio of 1:2.0 at 4°C for 30 min supplemented with 1 mM MgCl₂. The mixture was subsequently subjected to size-exclusion chromatography (Superose 6 increase 10/300, GE Healthcare) with buffer C to remove excess crRNA. Purity of the proteins was monitored at all stages of the purification process using SDS–PAGE (polyacrylamide gel electrophoresis) and visualized by Coomassie blue staining. crRNA was monitored using 10% denaturing TBE-urea PAGE and visualized by ethidium bromide staining.

Crystallization, data collection, structure determination and refinement. Crystals of the LbCpf1–crRNA complex were generated by mixing the protein complex with an equal amount of well solution (2 μ l) by the hanging-drop vapour-diffusion method. Crystals grew to their maximum size in five days in the solution containing 0.16 M Magnesium acetate and 18% (w/v) Polyethylene glycol (PEG) 3,350 at 20°C.

Before data collection, the crystals were transferred into cryo-protectant buffer (the crystallization buffer containing 20% (w/v) glycerol) and flash-cooled in liquid nitrogen. Diffraction data were collected at the Shanghai Synchrotron Radiation Facility (SSRF) at beam line BL17U1 using a CCD detector. The crystals belonged to space group C2 with one complex per asymmetric unit. For data collection, the crystals were equilibrated in a cryoprotectant buffer containing reservoir buffer plus 20% (v/v) glycerol. The data were processed using HKL2000 (ref. 31). Initial phases were obtained with a SeMet-crystal diffracting to 2.77 Å by the Se single-wavelength anomalous dispersion method using AutoSol³². The phases were then extended to the 2.38 Å data set collected from another SeMet-crystal. The electron density calculated to 2.38 Å was sufficient for model building with the program COOT³³. The built model was refined by the program PHENIX³⁴. The finally refined model contained residues 1–132, 135–280, 292–1078 and 1085–1228 of LbCpf1 and one crRNA molecule. The structure figures were prepared using PYMOL³⁵.

In vitro transcription and purification of crRNA. The crRNAs were transcribed *in vitro* using T7 polymerase and purified using denaturing PAGE using the following protocol. Transcription templates (dsDNA) were generated by PCR. Large scale transcription reactions (20 ml) were conducted in buffer containing 0.1 M HEPES-K, pH 7.9, 12 mM MgCl₂, 30 mM DTT, 2 mM Spermidine, 2 mM each NTP, 80 μ g ml⁻¹ home-made T7 polymerase and 300 nM transcription template. The reactions were performed at 37°C for 4–6 h and stopped by addition of 70% ethanol. The crRNA-containing pellets were then resuspended and purified by gel electrophoresis on a 10% denaturing (8 M urea) polyacrylamide gel. crRNA bands in the gel were excised and recovered with Elutrap System followed by ethanol precipitation. crRNAs were resuspended in diethyl pyrocarbonate H₂O and stored at –80°C.

Limited proteolysis assay. We prepared 6 μ M of LbCpf1, LbCpf1–crRNA (molar ratio, 1:2.0) and LbCpf1–crRNA–dsDNA (molar ratio, 1:2.0:2.0) in buffer

C containing 2 mM MgCl₂ and incubated at room temperature (20°C) for 30 min. Then the resulting samples were incubated with 7 ng μ l⁻¹ trypsin and incubated on ice for 40 min. The reactions were stopped by adding 2 \times SDS gel-loading buffer and 95°C quenched for 3 min. Samples were applied to 12% SDS–PAGE and visualized by Coomassie blue G-250 staining.

Microscale thermophoresis assay (MST). The affinity of the purified LbCpf1 protein with RNA was calculated using Monolith NT. 115 (NanoTemper Technologies GmbH, Munich, Germany)³⁶. Proteins were labelled with NT-647-NHS fluorescent dye. An RNA with varying concentrations (from 0.03 nM to 25 μ M) was incubated with 20 nM of labelled LbCpf1 at room temperature for 15 min in buffer containing 20 mM Tris (pH 7.5) and 100 mM NaCl. The sample was loaded into NanoTemper hydrophilic treated capillaries. Measurements were performed at 24°C using 40% LED power and 60% MST power. All experiments were repeated three times for each measurement. Data analyses were carried out using NanoTemper analysis software.

Sedimentation velocity analytical ultracentrifugation. Sedimentation velocity was performed by an XL-I analytical ultracentrifuge (Beckman Coulter) equipped with an eight-cell An-50 Ti rotor for interaction analysis of LbCpf1 and LbCpf1–crRNA complex at 4°C. The OD₂₈₀ is about 0.8. Buffer containing 10 mM Tris-HCl, pH 8.0, 100 mM NaCl was used as the reference solution. All samples were applied at a speed of 40,000 r.p.m. Absorbance scans were taken at 280 nm at the intervals of 0.003 cm size in a radial direction. The different sedimentation coefficients, *c*(s), and molecular weight were calculated by SEDFIT V14.4f software.

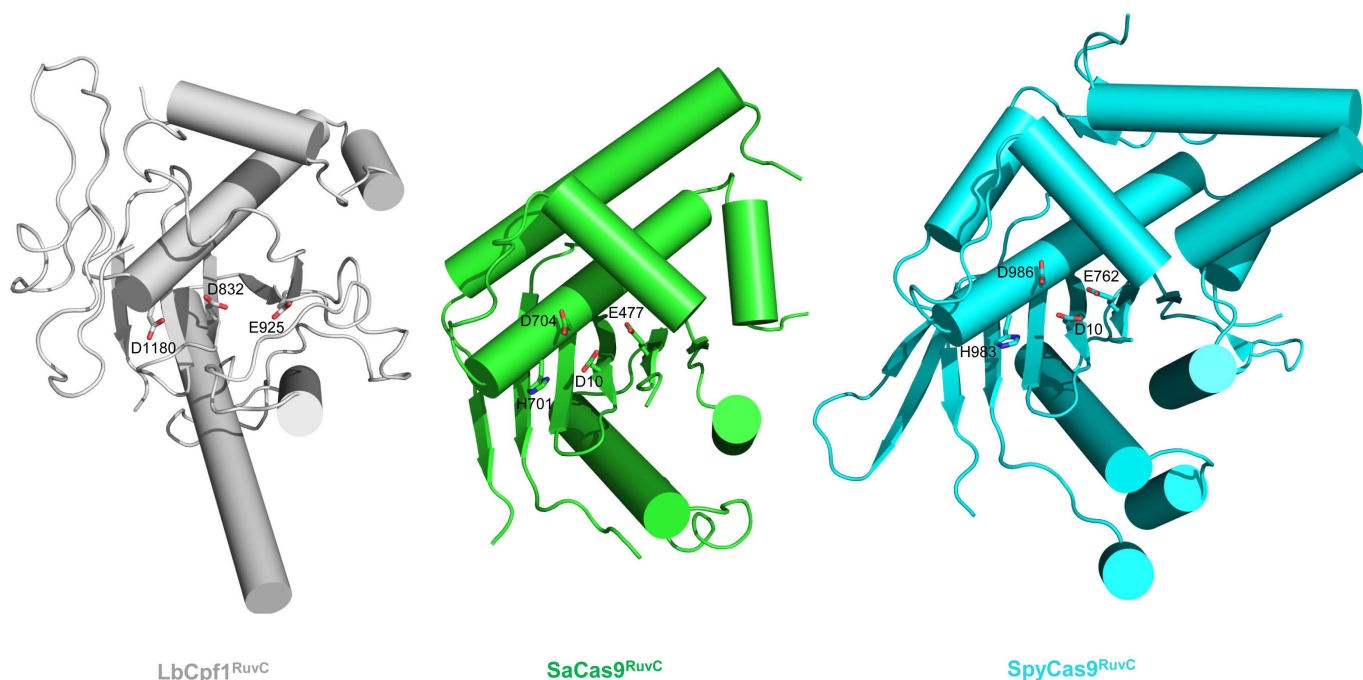
Negative staining electron microscopy. For negative staining electron microscopy, LbCpf1 and LbCpf1–crRNA complex were stained with 0.7% uranyl formate and 2% uranyl acetate, respectively. Samples of 4 μ l were applied to a carbon-coated copper grid and blotted with a piece of filter paper after a waiting time of 50–60 s. Grids were then washed with \sim 15 μ l stain solution for three times. 4 μ l stain solution was deposited on the grids and blotted after staining for 30 s, and left for air drying.

Data collection was performed on an FEI T12 microscope operated at 120 kV, and images were recorded using a 4 K \times 4 K charge-coupled device camera (UltraScan 4000, Gatan). Micrograph pre-processing and particle picking were performed with EMAN2 (ref. 37) and reference-free two-dimensional (2D) classification was performed with RELION³⁸.

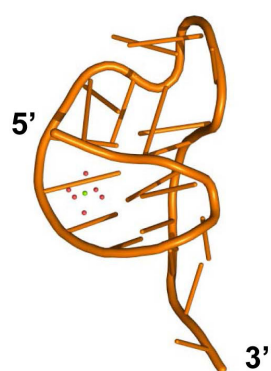
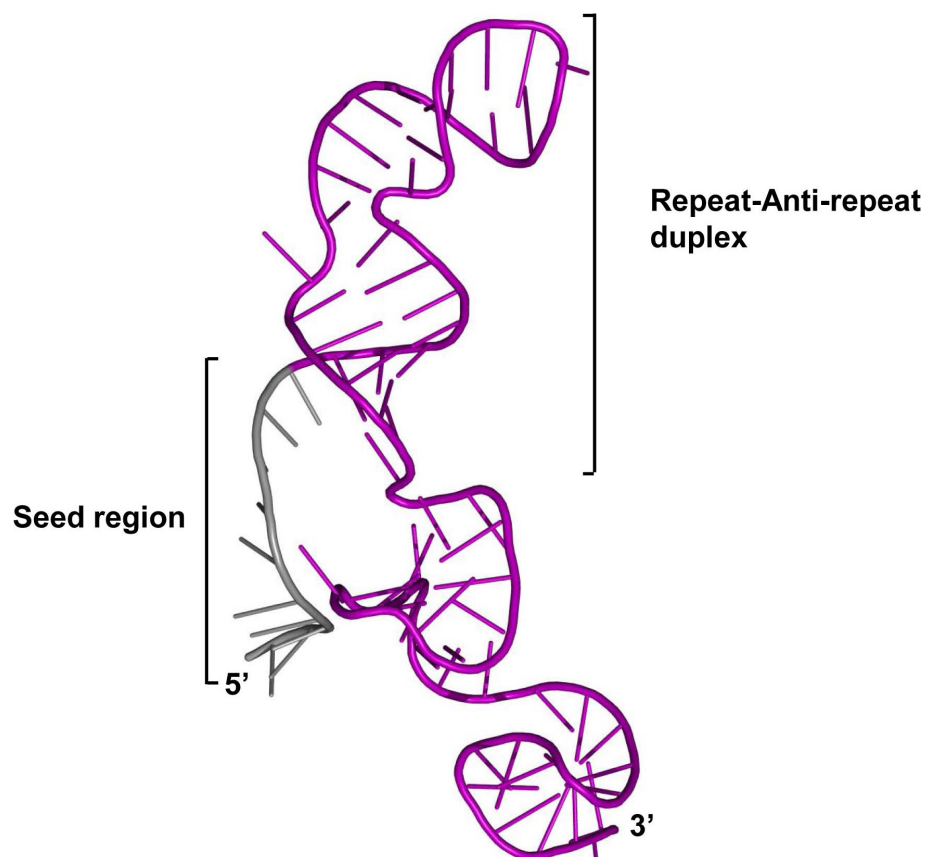
In vitro cleavage assay. *In vitro* dsDNA cleavage reactions were performed in a 50 μ l buffer system containing 3 μ g LbCpf1, 0.8 μ g crRNA and 1 μ g dsDNA. Target DNA sequence containing a protospacer target sequence and a 5'-TTA-3' PAM motif was cloned into pUC18 vector. To test RNA*-mediated inhibition of dsDNA cleavage by LbCpf1, molar ratios of RNA*:crRNA ranging from 0:1 to 64:1 were used. Cleavage reactions were conducted at 37°C for 10 min in cleavage buffer (50 mM Tris-HCl, pH 7.9, 10 mM MgCl₂, 100 mM NaCl, 5 mM DTT). Reactions were stopped by adding 2 \times TBE-urea gel loading buffer and 100°C quenching for 3 min. Cleavage products were run on TBE-urea 6% PAGE and visualized by EB staining.

GST pull-down assay. Purified mutant GST–LbCpf1 protein (150 μ g) was incubated with purified crRNA and dsDNA oligos (molar ratio, 1:2:3) at room temperature for 15 min. 50 μ l GS4B resin was added into each reaction system and incubated at 4°C for 20 min. After washing three times with buffer containing 10 mM Tris-HCl (pH 8.0) and 150 mM NaCl, the reaction mixtures were run on denaturing TBE-urea 10% PAGE and visualized by ethidium bromide staining. The experiment was repeated three times.

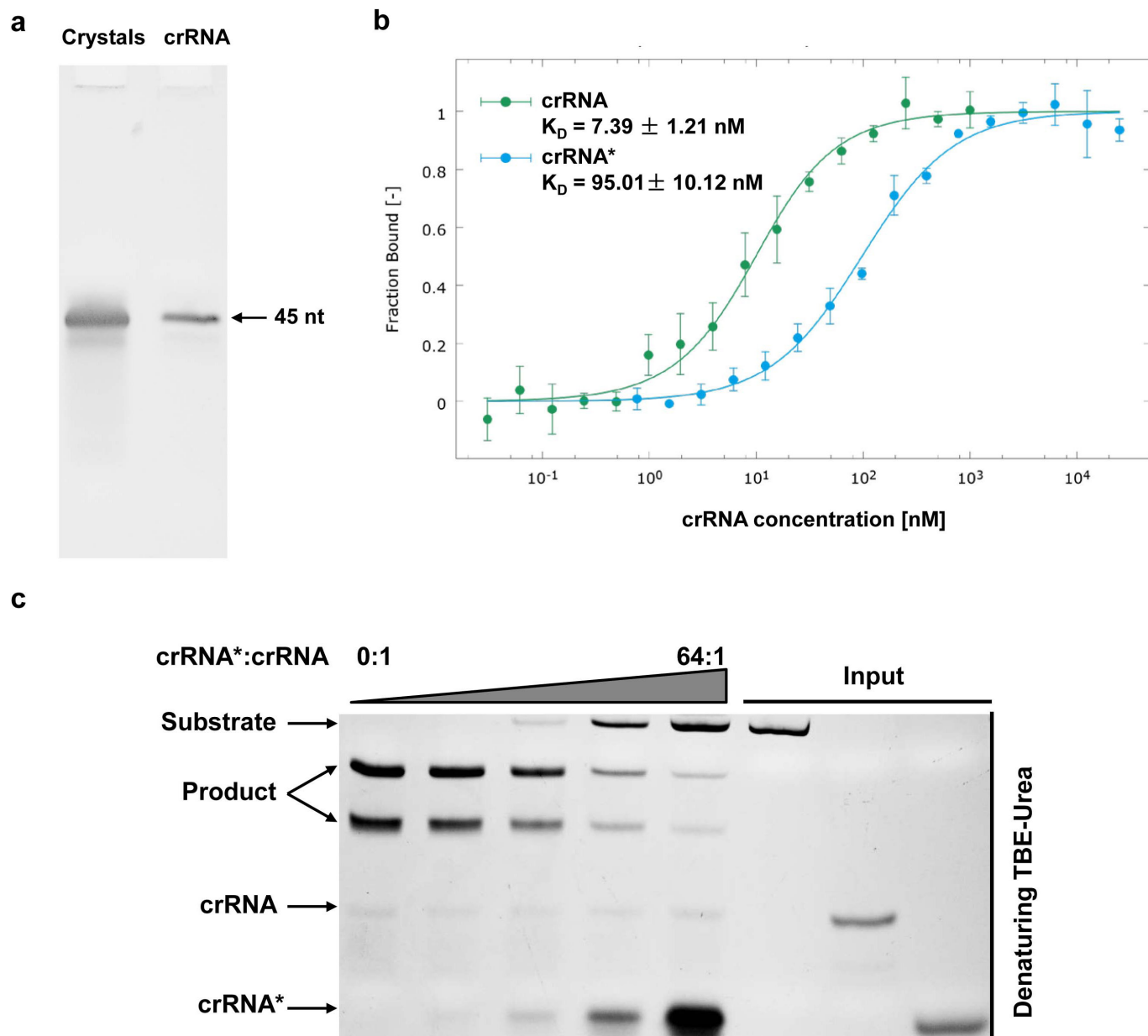
- Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
- McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
- Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
- Adams, P. D. *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58**, 1948–1954 (2002).
- DeLano, W. L. PyMOL Molecular Viewer (<http://www.pymol.org>) (2002).
- Wienken, C. J., Baaske, P., Rothbauer, U. & Braun, D. Protein-binding assays in biological liquids using microscale thermophoresis. *Nat. Commun.* **1**, 100 (2010).
- Tang, G. *et al.* EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157**, 38–46 (2007).
- Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).



Extended Data Figure 1 | Structural comparison of the RuvC domains of LbCpf1, SaCas9 and SpyCas9. Structural superposition of LbCpf1^{RuvC} with SaCas9^{RuvC} (PDB, 5CZZ) and SpyCas9^{RuvC} (PDB, 4UN3). The catalytic residues of the three RuvC domains are labelled. LbCpf1^{RuvC}, SaCas9^{RuvC} and SpyCas9^{RuvC} domains are coloured in grey, green and cyan, respectively.

LbCpf1-bound crRNA**SpyCas9-bound sgRNA**

Extended Data Figure 2 | Structural comparison of the LbCpf1-bound crRNA and the SpyCas9-bound sgRNA. Shown are the structures of LbCpf1-bound crRNA (left) and the SpyCas9-bound sgRNA (PDB, 4UN3) (right).



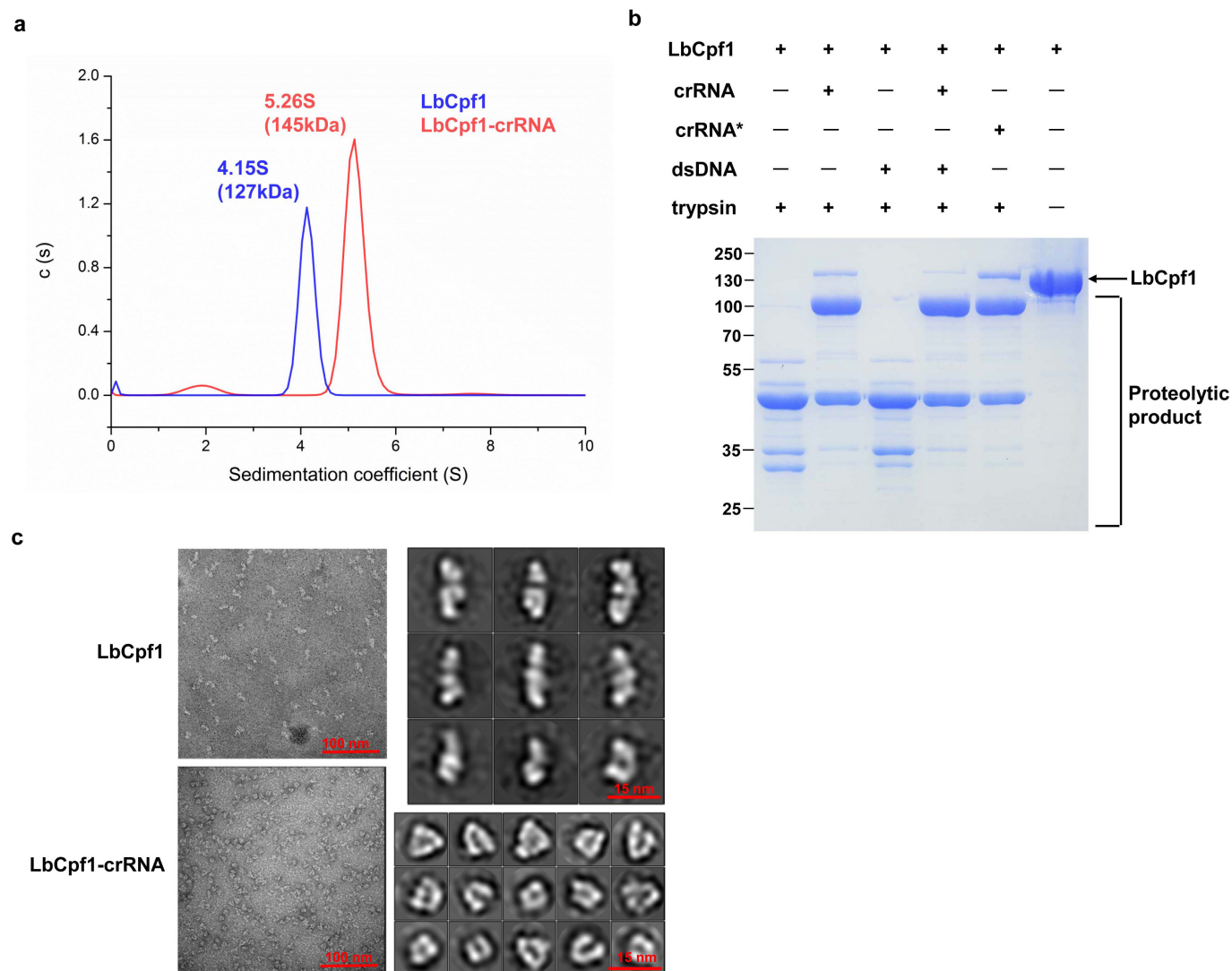
Extended Data Figure 3 | The direct repeat sequence of crRNA binds LbCpf1. **a**, crRNA in the crystal remains intact. Crystals of the LbCpf1–crRNA complex were collected and the integrity of crRNA was checked by denaturing TBE-urea (10%) polyacrylamide gel electrophoresis and stained by ethidium bromide. **b**, A crRNA lacking the guide sequence (crRNA*) binds LbCpf1. Data shown here are representative of three independent microscale thermophoresis experiments and the errors were calculated as standard deviation. **c**, crRNA* inhibits crRNA-guided LbCpf1 endonuclease activity *in vitro*. 0.8 μ g crRNA and 3 μ g purified LbCpf1 protein were mixed with varying amount of crRNA*. 1 μ g dsDNA was then added to the mixture that

was pre-incubated at 37 °C for 10 min. The nucleotides were analysed by running the mixture on TBE-urea polyacrylamide gels (10%) and visualized by ethidium bromide staining. Non-target strand sequence: 5'-TCGGTGC GGGCCTCTTCGCTATTACGCCAGCTGGCGA AAGGGGATGTGCTGCAAGGCGATTAAGTTGGGTAACGCCAGGG TTTTCCCAGTCACGACGTTGTAAAAACGACGCCAGTGCCAAGCTTG CATGCCTGCAGGTCGACTCTAGAGGATCCTTTAGAGAAGTCATTT AATAAGGCCACTGTAAAAAGCTTGGCGTAATCAGAATTCGTAAT CATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAA TTCCACACAACATACGAGCCGGAAGCATAAA-3'.

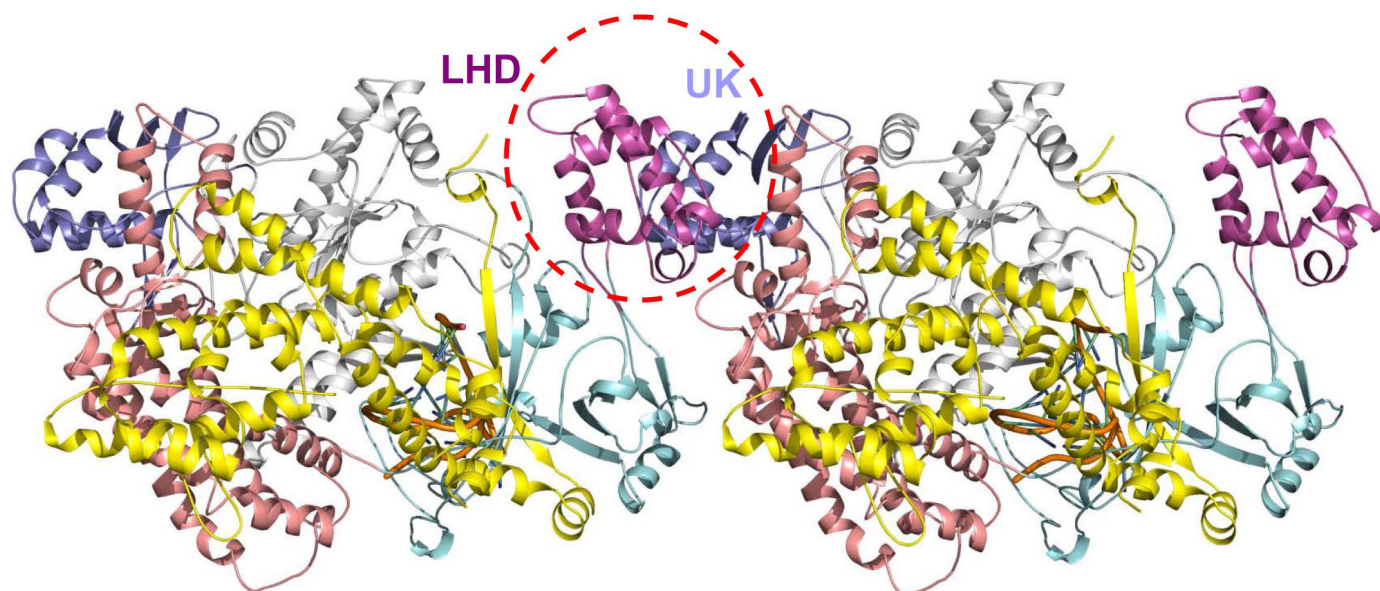


Extended Data Figure 4 | Sequence alignment of Cpf1 proteins from different species. Sequence alignment of Cpf1 proteins from different species. Conserved and similar residues are highlighted with red and yellow grounds respectively. Residues of LbCpf1 involved in crRNA

interaction are indicated with slate solid dots at bottom. α-helices and β-strands are shown as curly and arrow symbols, respectively. Protein domains identified in the structure are indicated.

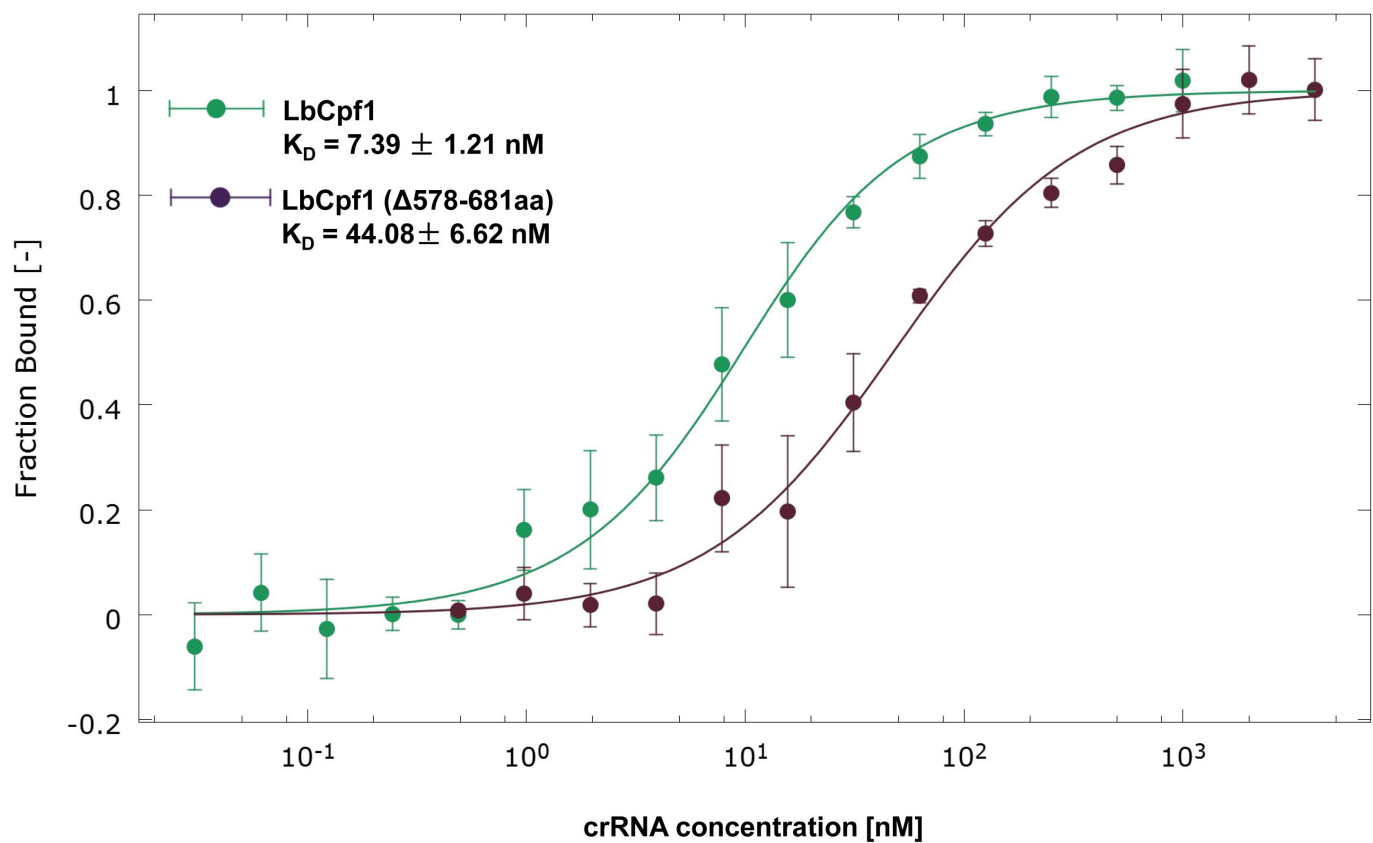


the absence or presence of crRNA or crRNA* for 30 min. The samples were then subjected to SDS–PAGE analysis. crRNA*, crRNA with the guide sequence deleted. **c**, Negative staining electron microscopy analysis of LbCpf1 and LbCpf1–crRNA complexes. Left, representative raw micrographs of negative-stained LbCpf1 (top) and LbCpf1–crRNA complex (bottom) samples. Right, representative 2D class averages of negatively stained particles of LbCpf1 (top) and LbCpf1–crRNA complex (bottom) samples.

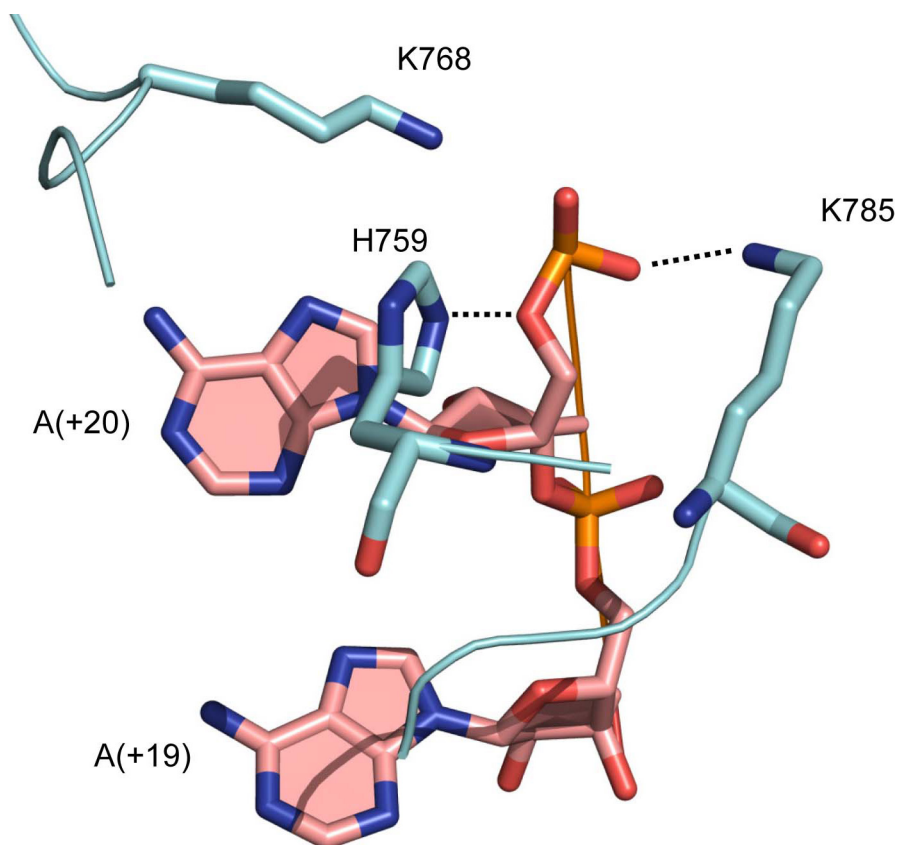


Extended Data Figure 6 | LHD is stabilized by interaction with LbCpf1 from a different asymmetric unit. LHD is involved in crystal packing in the LbCpf1–crRNA crystals. Shown in the figure are the LbCpf1–crRNA structures from two neighbouring asymmetric units. The LHD and UK

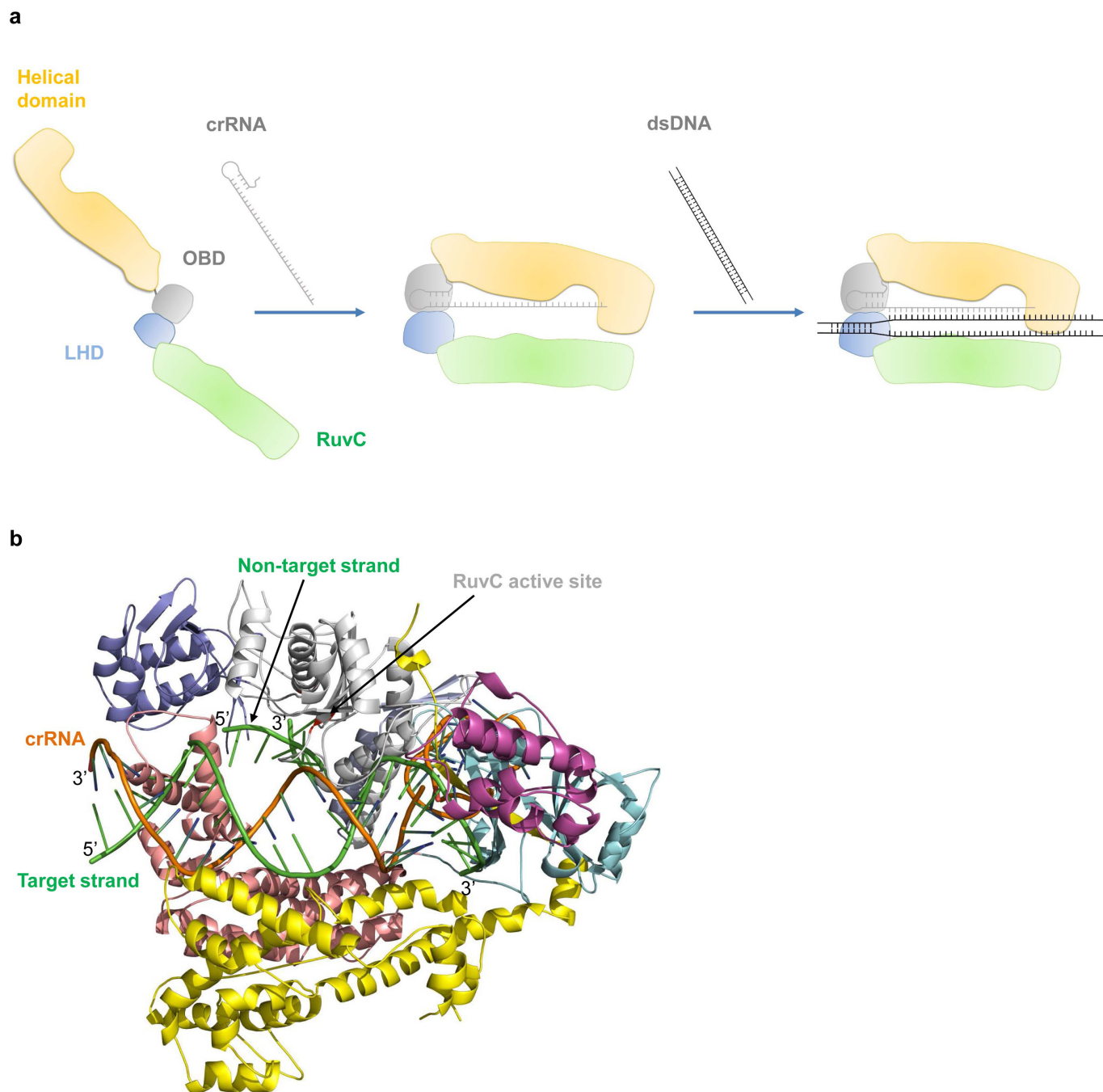
domains of LbCpf1 are shown in magenta and slate, respectively. Packing between the LHD from one asymmetric unit and the UK domain from a neighbouring asymmetric unit is marked with the red dashed circle.



Extended Data Figure 7 | A LHD-truncated LbCpf1 protein retains high binding affinity with crRNA. Data shown are representatives of three independent microscale thermophoresis experiments; error bars, s.d.



Extended Data Figure 8 | The interactions of 5' end of crRNA with LbCpf1. Detailed interactions of sugar-phosphate backbone of crRNA with LbCpf1 OBD. The residues from OBD responsible for LbCpf1 crRNA cleavage activity are labelled and shown in aquamarine.



Extended Data Figure 9 | A model of crRNA-guided LbCpf1 activation.
a, A model of LbCpf1 activation triggered by crRNA. The apo state LbCpf1 is maintained in an expended conformation. LbCpf1 is switched into a substrate-binding state through structural rearrangement trigger by crRNA binding to the OBD of LbCpf1. Then substrate DNA binds to

LbCpf1 with the involvement of the LHD and base pairs with the LbCpf1-bound crRNA, resulting in endonuclease cleavage of the dsDNA.
b, A model of non-target and target DNA (shown in green) bound to LbCpf1. Individual LbCpf1 domains are coloured according to the scheme in **a**, crRNA is shown in cartoon and coloured in orange.

Extended Data Table 1 | Data collection, phasing and refinement statistics

Data Collection Statistics

Data set	Se_LbCpf1-crRNA_1	Se_LbCpf1-crRNA_2
Beam Line	BL19U, SSRF	BL17U, SSRF
Space Group	C2	C2
Wavelength (Å)	0.97776	0.9794
Number of Reflections	65,357(1,285)	577,628(3,453)
Cell dimensions		
<i>a</i> , <i>b</i> , <i>c</i> (Å)	168.35, 83.91, 124.09	168.21, 82.97, 123.65
α , β , γ (°)	90, 106.72, 90	90, 106.73, 90
Resolution (Å)	41.51-2.38 (2.42-2.38)	50-2.77 (2.82-2.77)
R_{sym} (%)	5.8(64)	7.5(69)
$I/\sigma(I)$	20.67(1.4)	15.8(1.4)
Completeness (%)	99.6%(99.7%)	99.5% (99.4%)
Redundancy	3.8(3.8)	2.7(2.7)

Refinement Statistics

Resolution (Å)	41.5-2.38 (2.42-2.38)
No. Reflections	64,262
Completeness (%)	97.83%
Rwork/Rfree %	20.5(30.3)/26.1(34.6)
R.m.s.d	
Bond Lengths (Å)	0.009
Bond Angles (°)	1.388
Ramachandran (%)	
Preferred region	92.09
Allowed region	6
Outliers	1.92

Highest resolution shell is shown in parenthesis

Crystal structure of the human σ_1 receptor

Hayden R. Schmidt^{1*}, Sanduo Zheng^{1*}, Esin Gurbinar¹, Antoine Koehl², Aashish Manglik² & Andrew C. Kruse¹

The human σ_1 receptor is an enigmatic endoplasmic-reticulum-resident transmembrane protein implicated in a variety of disorders including depression, drug addiction, and neuropathic pain¹. Recently, an additional connection to amyotrophic lateral sclerosis has emerged from studies of human genetics and mouse models². Unlike many transmembrane receptors that belong to large, extensively studied families such as G-protein-coupled receptors or ligand-gated ion channels, the σ_1 receptor is an evolutionary isolate with no discernible similarity to any other human protein. Despite its increasingly clear importance in human physiology and disease, the molecular architecture of the σ_1 receptor and its regulation by drug-like compounds remain poorly defined. Here we report crystal structures of the human σ_1 receptor in complex with two chemically divergent ligands, PD144418 and 4-IBP. The structures reveal a trimeric architecture with a single transmembrane domain in each protomer. The carboxy-terminal domain of the receptor shows an extensive flat, hydrophobic membrane-proximal surface, suggesting an intimate association with the cytosolic surface of the endoplasmic reticulum membrane in cells. This domain includes a cupin-like β -barrel with the ligand-binding site buried at its centre. This large, hydrophobic ligand-binding cavity shows remarkable plasticity in ligand recognition, binding the two ligands in similar positions despite dissimilar chemical structures. Taken together, these results reveal the overall architecture, oligomerization state, and molecular basis for ligand recognition by this important but poorly understood protein.

The development of radiolabelled opiates in the 1960s and 1970s led to the discovery that the effects of these drugs are mediated by specific receptor sites with discrete pharmacological properties³. These receptors were divided into four classes based on their ligand-binding properties and tissue distribution, leading to the concept of μ (morphine), δ (vas deferens), κ (ketazocine), and σ (SKF-10047) opioid receptor subtypes. Pharmacological studies suggested μ , δ , and κ receptors were closely related to one another, while the σ receptor was shown to be distinct. Unlike canonical opioid receptors, the σ_1 receptor shows negligible affinity for naloxone and naltrexone. In addition, it exhibits a marked preference for the (+)-enantiomers of benzomorphan drugs while canonical opioid receptors bind with high affinity only to the (–)-enantiomers⁴. In 1996, the molecular cloning of the σ_1 receptor confirmed definitively that the receptor is dissimilar in sequence from the true opioid receptors⁵. The σ_1 receptor plays a key role in human physiology, and has been shown to modulate a variety of diseases of the cardiovascular and nervous system⁶. Of particular note, a point mutation in this receptor was identified as a cause of juvenile-onset amyotrophic lateral sclerosis in humans⁷, and mouse studies further support a role in the progression of this disease⁸. Other important research has suggested a role for the σ_1 receptor as an endoplasmic-reticulum chaperone protein and regulator of calcium signalling⁹, and it has been reported to regulate the activity of various ion channels¹⁰ and G-protein-coupled receptors¹¹.

Despite the increasingly apparent importance of the σ_1 receptor in human physiology, remarkably little is known about its structure and the

details of its function at the molecular level. Even the overall topology of the receptor has remained in doubt, with single-pass¹² and two-pass¹³ transmembrane architectures proposed. To address the gap in structural information surrounding the σ_1 receptor, we undertook biochemical and crystallographic studies to elucidate its structure in complex with two distinct ligands. A receptor construct bearing an amino (N)-terminal Flag tag was expressed in *Sf9* insect cells and purified in detergent (Extended Data Fig. 1). Using lipidic cubic phase crystallization we obtained crystals and used experimental phasing with tantalum bromide clusters to solve a 2.5 Å resolution structure of the σ_1 receptor bound to PD144418, a high-affinity and selective σ_1 antagonist^{14,15}. A similar approach also enabled structure determination for σ_1 receptor bound to a second ligand¹⁶, 4-IBP, at 3.2 Å resolution (Extended Data Table 1 and Extended Data Fig. 2). 4-IBP has an incompletely understood efficacy profile, with functional properties suggestive of either agonist or inverse agonist activity¹⁷.

The overall structure of the σ_1 receptor reveals a trimeric organization with a threefold non-crystallographic symmetry axis normal to the membrane plane (Fig. 1a). The receptor contains only a single transmembrane domain for each protomer, contrary to the prevailing models of a two-pass transmembrane architecture. The carboxy (C)-terminal membrane-adjacent domains mediate the trimeric structure of the receptor, packing closely together with an interface of $\sim 9,300$ Å² between each adjacent pair of protomers. In contrast, the three transmembrane helices are widely separated from one another, located at each corner of the triangular trimer where they mediate lattice contacts (Extended Data Fig. 3). The membrane-proximal side of the cytosolic domains is an extremely flat hydrophobic surface, which is probably embedded within the membrane plane (Fig. 1b and Extended Data Fig. 4). The cytosolic domain of each of the three protomers shows a β -barrel fold with the ligand at its centre, flanked by four α -helices (Fig. 2). The ligand-binding domain is highly conserved in sequence across species, as is the intermolecular interface among the three protomers (Extended Data Figs 5 and 6). The overall fold of the β -barrel ligand-binding region closely resembles that of cupin family proteins, most of which are oligomeric bacterial enzymes (Extended Data Table 2). While there is no obvious functional similarity between such proteins and the σ_1 receptor, the enzyme catalytic sites are generally synonymous with the ligand-binding site of the σ_1 receptor, suggesting the σ_1 receptor may represent a repurposed enzyme in which the catalytic site inside the β -barrel has been co-opted as a ligand-binding site.

Recently, a mutation in the σ_1 receptor, E102Q, was identified as a cause of inherited juvenile-onset amyotrophic lateral sclerosis in a family in eastern Saudi Arabia⁷. Cell biological experiments have shown that this receptor mutant is prone to aggregation, leading to mislocalization of TDP43 and consequent cytotoxicity¹⁸. The structures reported here offer an explanation for the phenotype of this mutant. The highly conserved Glu102 is deeply buried, with its carboxyl oxygen atoms each accepting a hydrogen bond from the backbone amides of Val36 and Phe37, which are part of a structured tether between the transmembrane domain and cytosolic domain (Fig. 2c). Mutation of Glu to Gln would block this interaction, converting one of the two favourable

¹Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA. ²Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, California 94305, USA.

*These authors contributed equally to this work.

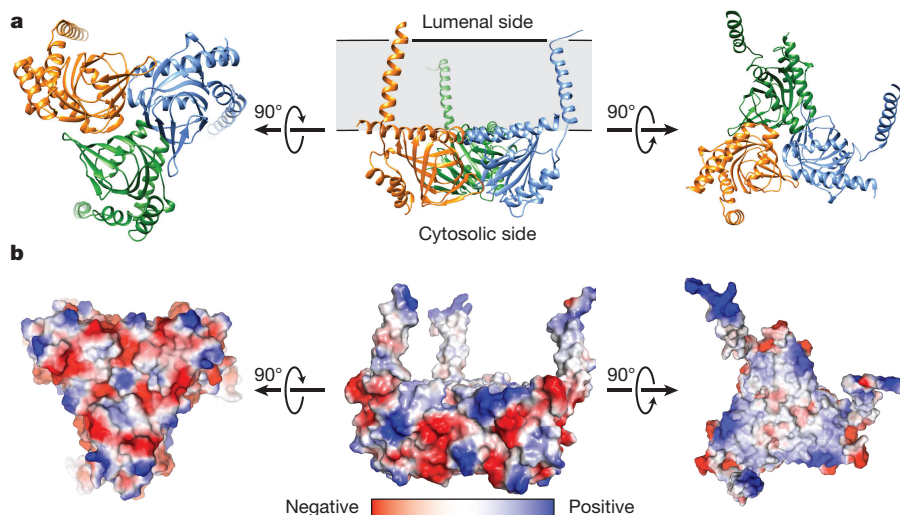


Figure 1 | Overall structure of the σ_1 receptor. **a**, Viewed perpendicular to the membrane plane, the σ_1 receptor shows a triangular structure comprising three tightly associated protomers, each with a single transmembrane domain at a corner of the oligomeric triangle. From the side, the receptor reveals a flat membrane-associated surface. The location

of the membrane plane is shown in grey, on the basis of PPM server³⁰ prediction. **b**, Colouring by electrostatic potential reveals a polar cytosolic surface (left side), and a non-polar membrane-interacting surface flanked by positive charges, suggesting it is partly buried in the membrane.

hydrogen bond interactions into an energetically unfavourable juxtaposition of hydrogen bond donors, accounting for the previously observed receptor destabilization.

One of the most intriguing features of the σ_1 receptor is the remarkable diversity of the ligands to which it binds. These include a multitude

of biologically active compounds targeted at other receptors, such as dextromethorphan (inhibition constant $K_i = 200$ nM)¹⁹, haloperidol ($K_i = 1.1$ nM), fluoxetine ($K_i = 1.9$ μ M), quetiapine ($K_i = 220$ nM), clemastine ($K_i = 67$ nM), and chloroquine ($K_i = 109$ nM), among many others (affinities are from the PDSP K_i Database²⁰ unless otherwise noted). These ligands are diverse in chemical structure, sharing few common features with the exception of a cationic amine and at least one aromatic ring. To understand the molecular basis for this ligand-binding promiscuity, we performed additional crystallization and structure determination experiments with receptor bound to the ligand 4-IBP. The two ligands, 4-IBP and PD144418, were selected in part on the basis of their divergence in chemical structure, with a Tanimoto similarity coefficient of 0.235, indicating no substantial structural similarity. Nonetheless, it should be noted that both compounds are positively charged, elongated molecules with substantial hydrophobic character—all common features among σ_1 receptor ligands.

In comparing the two structures, very little deviation is seen in receptor conformation, and the all-atom root mean squared deviation for the two structures is 0.4 Å. The two ligands bind in similar positions (Extended Data Fig. 7), in each case interacting with the receptor through a charge–charge interaction with the highly conserved Glu172, consistent with previous mutagenesis experiments identifying this residue as essential for ligand binding²¹. A second essential acidic residue, Asp126, forms a 2.7 Å hydrogen bond with Glu172, indicating it is probably protonated at least when ligands are bound. With the exception of these two amino acids, the binding pocket overall is very hydrophobic and its interior is completely occluded from solvent (Fig. 3a). Other residues in the binding site include Val84, Trp89, Met93, Leu95, Leu105, Phe107, Ile124, Trp164, and Leu182, which interact with hydrophobic regions on the bound ligands, and Tyr103, which engages in an aromatic stacking interaction in both structures (Fig. 3b, c). In addition, Tyr103 makes a hydrogen bond to Glu172, accounting for a fivefold reduction in binding affinity to (+)-pentazocine in a Y103F mutant²².

Given the highly occluded structure of the binding pocket, it remains unclear how ligands enter and exit this site. Two possibilities are apparent: ligands could enter and exit through a gap between the two membrane-adjacent helices, directly into/out of the plasma membrane, or they could access the binding site through the cytosolic surface, passing through a polar region occluded by Gln135, Glu158, and His154. Since both potential points of entry/egress are in a closed conformation in

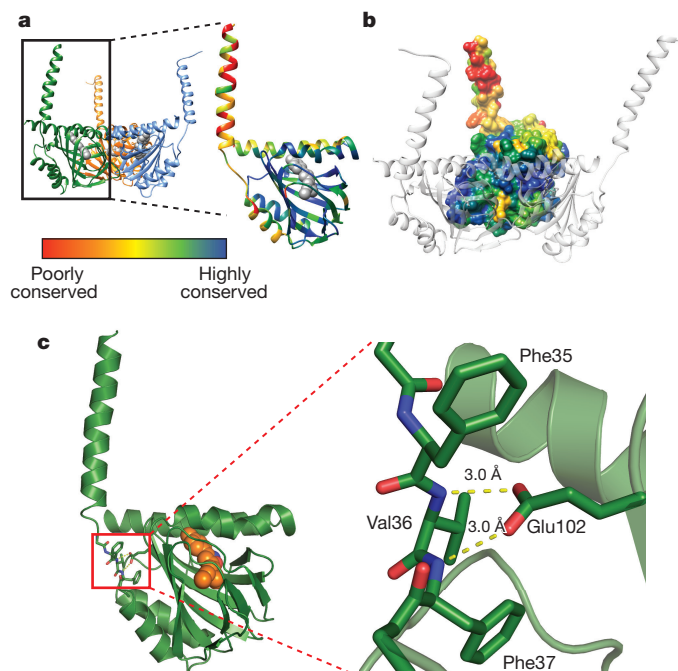


Figure 2 | Structure of the σ_1 protomer. **a**, The receptor shows a cupin-like β -barrel fold flanked by four α -helices with the ligand (grey) bound at the centre of the cupin domain. The receptor is coloured by sequence conservation, revealing a high degree of conservation in the ligand-binding domain, and relatively lower conservation of the transmembrane helices, which may simply act to tether the receptor to the membrane. **b**, The intermolecular interface among protomers of the receptor trimer is likewise highly conserved. **c**, Glu102 forms a pair of hydrogen bonds (yellow dashed lines) with backbone amide nitrogen atoms, providing a structural explanation for receptor destabilization due to the E102Q mutation associated with amyotrophic lateral sclerosis.

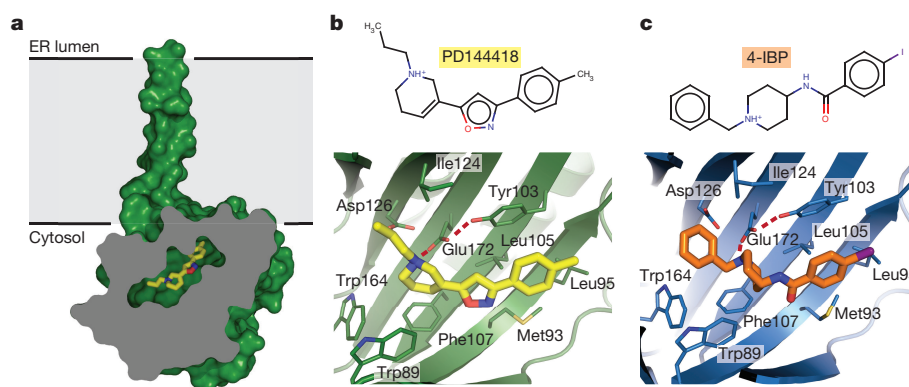


Figure 3 | Ligand recognition. **a**, Cross-section view of the receptor bound to PD144418, showing the deeply buried antagonist and occlusion of the binding pocket from solvent. The ligand is shown in yellow sticks. **b**, View of PD144418 binding pose, showing charge–charge interaction

with Glu172 (red dotted line) and extensive hydrophobic contacts with other binding pocket residues. A hydrogen bond between Glu172 and Tyr103 is also shown as a red dotted line. **c**, Corresponding structure of the 4-IBP binding pose.

the current structures, some degree of conformational plasticity must exist to account for reversible ligand binding. Notably, the occluded structure of the binding site accounts for the very slow ligand-binding kinetics typically seen with σ_1 receptor, and the resulting requirement to use elevated temperatures or very long incubation times to reach equilibrium in radioligand binding assays²³.

A key unanswered question surrounding σ_1 receptor function is the molecular basis of ligand efficacy. The classification of σ_1 ligands as agonists and antagonists is largely based on whole-animal physiology, with agonists defined as ligands that induce hyperlocomotion or other physiological responses through binding to σ_1 , while antagonists are σ_1 ligands that block or blunt this response^{24,25}. In addition, antagonists show similar functional effects to receptor knockdown, suggesting they indeed operate through blockade of σ_1 activity²⁶. The relationship between ligand binding to σ_1 receptor and the subsequent biological response remains only partly understood²⁷. However, a recent study²⁸ offered biochemical evidence for ligand-mediated changes in σ_1 receptor oligomerization state. Subsequent FRET studies in cells have revealed similar results, showing in addition that antagonists stabilize high molecular mass oligomers, while agonists favour dissociation of these complexes²⁹.

To better understand σ_1 receptor oligomerization, we performed size-exclusion chromatography with multi-angle light scattering (SEC–MALS) experiments as well as native polyacrylamide gel electrophoresis (PAGE) analysis. Samples in SEC–MALS showed a single sharp peak of protein, but light scattering and refractive index analysis revealed that this peak comprises protein species ranging in molecular mass from at least 140 kDa (excluding detergent mass) to about 400 kDa. This suggests the presence of oligomers ranging in size from hexamers to as large as 15-mers (Extended Data Fig. 8a, b). Native PAGE experiments showed similar results, again revealing a polydisperse mixture of high molecular mass oligomers (Extended Data Fig. 8c). These experiments in pure detergents showed little difference between agonist- and antagonist-bound receptor. In contrast, size exclusion in a mixed micelle of maltose neopentyl glycol detergent with cholesterol hemisuccinate showed modest differences in SEC profile (Extended Data Fig. 8d), with agonists partly disrupting high-order oligomers. It is important to note that our results in detergent may not fully recapitulate receptor behaviour *in vivo*; however, taken together with the previously reported biochemical and cellular studies, these data suggest oligomerization is a key functional property of the σ_1 receptor and may be linked to ligand efficacy.

In summary, the results presented here show for the first time the overall molecular structure of the σ_1 receptor, an important but poorly understood human transmembrane receptor. The structure reveals the basis for receptor oligomerization and ligand binding, and moreover

shows an unexpected single-pass transmembrane topology. These results now offer a solid foundation for the development of future biochemical and biophysical studies towards understanding the σ_1 receptor at the molecular level.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 16 October 2015; accepted 2 February 2016.

Published online 4 April 2016.

- Cobos, E. J., Entrena, J. M., Nieto, F. R., Cendán, C. M. & Del Pozo, E. Pharmacology and therapeutic potential of sigma₁ receptor ligands. *Curr. Neuropharmacol.* **6**, 344–366 (2008).
- Mavlyutov, T. A., Guo, L. W., Epstein, M. L. & Ruoho, A. E. Role of the sigma-1 receptor in amyotrophic lateral sclerosis (ALS). *J. Pharmacol. Sci.* **127**, 10–16 (2015).
- Brownstein, M. J. A brief history of opiates, opioid peptides, and opioid receptors. *Proc. Natl Acad. Sci. USA* **90**, 5391–5393 (1993).
- Largent, B. L., Wikström, H., Gundlach, A. L. & Snyder, S. H. Structural determinants of sigma receptor affinity. *Mol. Pharmacol.* **32**, 772–784 (1987).
- Hanner, M. *et al.* Purification, molecular cloning, and expression of the mammalian sigma₁-binding site. *Proc. Natl Acad. Sci. USA* **93**, 8072–8077 (1996).
- Su, T. P., Hayashi, T., Maurice, T., Buch, S. & Ruoho, A. E. The sigma-1 receptor chaperone as an inter-organelle signaling modulator. *Trends Pharmacol. Sci.* **31**, 557–566 (2010).
- Al-Saif, A., Al-Mohanna, F. & Bohlega, S. A mutation in sigma-1 receptor causes juvenile amyotrophic lateral sclerosis. *Ann. Neurol.* **70**, 913–919 (2011).
- Mavlyutov, T. A. *et al.* Lack of sigma-1 receptor exacerbates ALS progression in mice. *Neuroscience* **240**, 129–134 (2013).
- Hayashi, T. & Su, T. P. Sigma-1 receptor chaperones at the ER-mitochondrion interface regulate Ca²⁺ signaling and cell survival. *Cell* **131**, 596–610 (2007).
- Wu, Z. & Bowen, W. D. Role of sigma-1 receptor C-terminal segment in inositol 1,4,5-trisphosphate receptor activation: constitutive enhancement of calcium signaling in MCF-7 tumor cells. *J. Biol. Chem.* **283**, 28198–28215 (2008).
- Kim, F. J. *et al.* σ_1 Receptor modulation of G-protein-coupled receptor signaling: potentiation of opioid transduction independent from receptor binding. *Mol. Pharmacol.* **77**, 695–703 (2010).
- Dussossoy, D. *et al.* Colocalization of sterol isomerase and sigma₁ receptor at endoplasmic reticulum and nuclear envelope level. *Eur. J. Biochem.* **263**, 377–386 (1999).
- Aydar, E., Palmer, C. P., Klyachko, V. A. & Jackson, M. B. The sigma receptor as a ligand-regulated auxiliary potassium channel subunit. *Neuron* **34**, 399–410 (2002).
- Akunne, H. C. *et al.* The pharmacology of the novel and selective sigma ligand, PD 144418. *Neuropharmacology* **36**, 51–62 (1997).
- Lever, J. R. *et al.* Relationship between cerebral sigma-1 receptor occupancy and attenuation of cocaine's motor stimulatory effects in mice by PD144418. *J. Pharmacol. Exp. Ther.* **351**, 153–163 (2014).
- John, C. S., Vilner, B. J. & Bowen, W. D. Synthesis and characterization of [¹²⁵I]-N-(N-benzylpiperidin-4-yl)-4-iodobenzamide, a new σ receptor radiopharmaceutical: high-affinity binding to MCF-7 breast tumor cells. *J. Med. Chem.* **37**, 1737–1739 (1994).
- Bermack, J. E. & Debonnel, G. Distinct modulatory roles of sigma receptor subtypes on glutamatergic responses in the dorsal hippocampus. *Synapse* **55**, 37–44 (2005).

18. Tagashira, H., Shinoda, Y., Shioda, N. & Fukunaga, K. Methyl pyruvate rescues mitochondrial damage caused by *SIGMAR1* mutation related to amyotrophic lateral sclerosis. *Biochim. Biophys. Acta* **1840**, 3320–3334 (2014).
19. Shin, E. J. *et al.* Dextromethorphan attenuates trimethyltin-induced neurotoxicity via σ_1 receptor activation in rats. *Neurochem. Int.* **50**, 791–799 (2007).
20. Roth, B. L., Lopez, E., Patel, S. & Kroeze, W. K. The multiplicity of serotonin receptors: uselessly diverse molecules or an embarrassment of riches? *Neuroscientist* **6**, 252–262 (2000).
21. Seth, P. *et al.* Expression pattern of the type 1 sigma receptor in the brain and identity of critical anionic amino acid residues in the ligand-binding domain of the receptor. *Biochim. Biophys. Acta* **1540**, 59–67 (2001).
22. Yamamoto, H. *et al.* Amino acid residues in the transmembrane domain of the type 1 sigma receptor critical for ligand binding. *FEBS Lett.* **445**, 19–22 (1999).
23. Kovács, K. J. & Larson, A. A. Up-regulation of [3 H]DTG but not [3 H](+)-pentazocine labeled sigma sites in mouse spinal cord by chronic morphine treatment. *Eur. J. Pharmacol.* **350**, 47–52 (1998).
24. Martin, W. R., Eades, C. G., Thompson, J. A., Huppler, R. E. & Gilbert, P. E. The effects of morphine- and nalorphine-like drugs in the nondependent and morphine-dependent chronic spinal dog. *J. Pharmacol. Exp. Ther.* **197**, 517–532 (1976).
25. Nguyen, L. *et al.* Role of sigma-1 receptors in neurodegenerative diseases. *J. Pharmacol. Sci.* **127**, 17–29 (2015).
26. Mei, J. & Pasternak, G. W. σ_1 Receptor modulation of opioid analgesia in the mouse. *J. Pharmacol. Exp. Ther.* **300**, 1070–1074 (2002).
27. Maurice, T. & Su, T. P. The pharmacology of sigma-1 receptors. *Pharmacol. Ther.* **124**, 195–206 (2009).
28. Gromek, K. A. *et al.* The oligomeric states of the purified sigma-1 receptor are stabilized by ligands. *J. Biol. Chem.* **289**, 20333–20344 (2014).
29. Mishra, A. K. *et al.* The sigma-1 receptors are present in monomeric and oligomeric forms in living cells in the presence and absence of ligands. *Biochem. J.* **466**, 263–271 (2015).
30. Lomize, M. A., Pogozheva, I. D., Joo, H., Mosberg, H. I. & Lomize, A. L. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res.* **40**, D370–D376 (2012).

Acknowledgements We thank beamline staff at the Advanced Photon Source for their technical assistance and support. We thank S. Blacklow and B. Zimmerman for discussions and input throughout the project, and K. Arnett for assistance with SEC–MALS experiments. Financial support for this work was provided by departmental startup funds. S.Z. is supported by a Merck Postdoctoral Fellowship in Biological Chemistry and Molecular Pharmacology, and H.R.S. is supported by the National Institutes of Health Cellular and Developmental Biology training grant at Harvard Medical School (T32GM007226).

Author Contributions Receptor purification and crystallization experiments were conducted by H.R.S., E.G., and A.C.K. X-ray data collection was performed by H.R.S., A.M., A.K., and A.C.K. Data processing, phase calculation, and structure refinement were performed jointly by H.R.S., S.Z., and A.C.K. SEC–MALS experiments were performed by A.C.K., radioligand binding by H.R.S., and native-PAGE by S.Z. Overall project design, molecular cloning, and pilot studies were conducted by A.M. and A.C.K.

Author Information Coordinates and structure factors for the σ_1 receptor bound to PD144418 and 4-IBP have been deposited in the Protein Data Bank under accession numbers 5HK1 and 5HK2, respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.C.K. (andrew_kruse@hms.harvard.edu).

METHODS

Expression and purification. The human σ_1 receptor was cloned into pFastBac1 with an N-terminal haemagglutinin signal sequence followed by a Flag epitope tag and a 3C protease cleavage site. After proteolytic digest to remove the Flag tag, the resulting protein is identical to the wild-type receptor with the exception of an N-terminal protease site scar, comprising the sequence 'GPGS'. This receptor construct was expressed in *Sf9* insect cells (Expression Systems) using the FastBac baculovirus system (ThermoFisher) according to the manufacturer's instructions. Infection was performed when cells reached a density of 4×10^6 cells per millilitre, and flasks were shaken at 27°C for 2 days before harvest.

Cells were harvested by centrifugation and frozen at -80°C until purification. For both PD144418 and 4-IBP-bound receptors, $1\ \mu\text{M}$ of ligand was added in all purification steps. After thawing frozen cell paste, cells were lysed by osmotic shock in 20 mM HEPES pH 7.5, 2 mM magnesium chloride, and 1:100,000 (v/v) benzamide nuclease (Sigma Aldrich). Lysed cells were centrifuged at 47,800g in a Sorvall RC 5C Plus centrifuge with an SS-34 rotor for 15 min. The receptor was then extracted using a glass dounce tissue grinder in a solubilization buffer containing 250 mM NaCl, 20 mM HEPES pH 7.5, 20% (v/v) glycerol, 1% (w/v) lauryl maltose neopentyl glycol (LMNG; Anatrace), and 0.1% (w/v) cholesterol hemisuccinate (CHS; Steraloids). Samples were stirred for 2 h at 4°C, then centrifuged as before for 20 min. Next, samples were filtered on a glass microfibre filter. The filtered supernatant containing solubilized receptor was supplemented with 2 mM calcium chloride and loaded by gravity flow onto 5 ml anti-Flag antibody affinity resin. The resin was washed extensively, first in 50 ml of buffer containing 100 mM NaCl, 20 mM HEPES pH 7.5, 2 mM calcium chloride, 0.2% glycerol, 0.1% LMNG, and 0.01% CHS, and then in 50 ml of buffer containing 100 mM NaCl, 20 mM HEPES pH 7.5, 2 mM calcium chloride, 0.02% glycerol, 0.01% LMNG, and 0.001% CHS. The receptor was eluted in the same buffer supplemented with 5 mM EDTA and 0.2 mg ml $^{-1}$ Flag peptide in lieu of calcium. 3C protease was added (1:100 w:w) and incubated with the receptor at 4°C overnight.

The receptor was further purified by SEC on a Sephadex S200 column (GE Healthcare) in buffer containing 0.01% LMNG, 0.001% CHS, 100 mM NaCl, 20 mM HEPES pH 7.5, and $1\ \mu\text{M}$ of ligand. The receptor was biochemically pure but consistently ran as a high molecular mass oligomer during SEC. After preparative SEC, the protein was concentrated to 20–30 mg ml $^{-1}$ and flash frozen with liquid nitrogen in aliquots of 8–9 μl . Samples were stored at -80°C until use for crystallography. Purity and monodispersity of crystallographic samples was evaluated by SDS-PAGE and analytical SEC, respectively (Extended Data Fig. 1).

Crystallography and data collection. Purified σ_1 receptor was reconstituted into lipidic cubic phase by mixing with a 10:1 (w:w) mix of monoolein (Hampton Research) with cholesterol (Sigma Aldrich) at a ratio of 1.5:1.0 lipid:protein by mass, using the coupled syringe reconstitution method³¹. All samples were mixed at least 100 times. The resulting phase was dispensed in 30–40 nl drops onto either a glass plate or a hanging drop cover, and overlaid with 600 nl of precipitant solution using a Gryphon LCP robot (Art Robbins Instruments). Crystals grew in precipitant solution containing 40–50% PEG 300, 220–250 mM LiSO₄, 0.1 M MES pH 6.5. Initial crystallization hits grew slowly, with crystals reaching full size over the course of 2–4 weeks. Crystals were harvested using mesh loops and stored in liquid nitrogen until data collection.

Data collection was performed at Advanced Photon Source GM/CA beamlines 23ID-B and 23ID-D (native data), and at NE-CAT beamline 24ID-C (tantalum bromide derivative). An initial grid raster with $80\ \mu\text{m} \times 30\ \mu\text{m}$ beam dimensions was performed using a 20 μm beam to locate crystals in the loop. Additional rasters were performed using a 10 μm beam diameter to optimally position the crystal for data collection. Data collection used a 10 μm beam and diffraction images were collected in 0.2–1° oscillations at a wavelength of 1.033 Å. For σ_1 bound to PD144418, a complete data set was obtained from a single crystal. For σ_1 bound to 4-IBP, a complete data set was the result of merging data from three crystals.

Experimental phasing and structure refinement. To obtain phases, crystals of σ_1 receptor bound to PD144418 were grown using a hanging-drop lipidic cubic phase methodology adapted from a previous report³². In brief, this entailed dispensing cubic phase drops onto a plastic cover film (Art Robbins Instruments) and overlaying with precipitant solution as described above. This film was then inverted over a matched plate with identical crystallization solutions to the precipitant surrounding the lipid drop. The resulting crystals could be soaked and resealed, unlike conventional glass sandwich lipidic cubic phase plates. Crystals prepared in this way were soaked with tantalum bromide clusters for approximately 12 h by adding crushed granules of tantalum bromide clusters to the edge of the well. The crystals were harvested and data collected as described above, but at a wavelength of 1.2548 Å.

Initial phases were obtained in SHARP³³ using single isomorphous replacement and anomalous scattering. Three transmembrane α -helices were identifiable in the

initial map, suggesting three molecules in the asymmetric unit with an unusual solvent content of $\sim 70\%$. Experimental phases were iteratively combined with model-derived phase to improve the electron density map through solvent flattening in SHARP. Model building was performed in Coot³⁴, and refinement was performed in phenix.refine³⁵. All three chains are highly similar in structure, with all-atom pairwise root mean squared deviation of cytosolic domains ranging from 0.22 Å to 0.26 Å, while the orientation of the transmembrane helix relative to the soluble domain varies among protomers.

Assignment of sequence register was straightforward and unambiguous owing to the relatively high resolution, almost completely ordered structure, and high-frequency bulky amino-acid side chains (σ_1 receptor is roughly 5% tryptophan). As a control for register assignment, the structure was built and register assigned in two independent ways. First it was manually built and register assigned by inspection of electron density. In parallel, sequence register was independently assigned automatically with phenix.autobuild, and results were confirmed to be identical throughout the entire polypeptide chain of each protomer. Representative composite omit map density is shown in Extended Data Fig. 2. Ligands were manually placed into $F_o - F_c$ difference maps (Extended Data Fig. 7). In the case of PD144418 the electron density was clear, and ligand position and pose were unambiguous. For 4-IBP, the pose was unambiguous owing to the high $F_o - F_c$ peak resulting from the ligand iodine atom. After refinement, structure quality was assessed using MolProbity³⁶, and figures were prepared in PyMOL³⁷ and UCSF Chimera³⁸. All crystallographic data processing, refinement, and analysis software was compiled and supported by the SBGrid Consortium³⁹.

Sequence and structure conservation analysis. Sequence conservation analysis in Fig. 2 was computed using the ConSurf server⁴⁰. In brief, a multiple sequence alignment of human σ_1 receptor to its closest 330 homologues was generated using a protein sequence BLAST search on the NCBI public database using the human wild-type σ_1 receptor protein sequence as query. These sequences were then used for ConSurf analysis, with conservation scores plotted using UCSF Chimera. Analysis of fold conservation was performed using the DALI server⁴¹ with the PD144418-bound structure of the σ_1 receptor as query. Structures with Z-scores in excess of 8 were selected for further analysis, with results summarized in Extended Data Table 2.

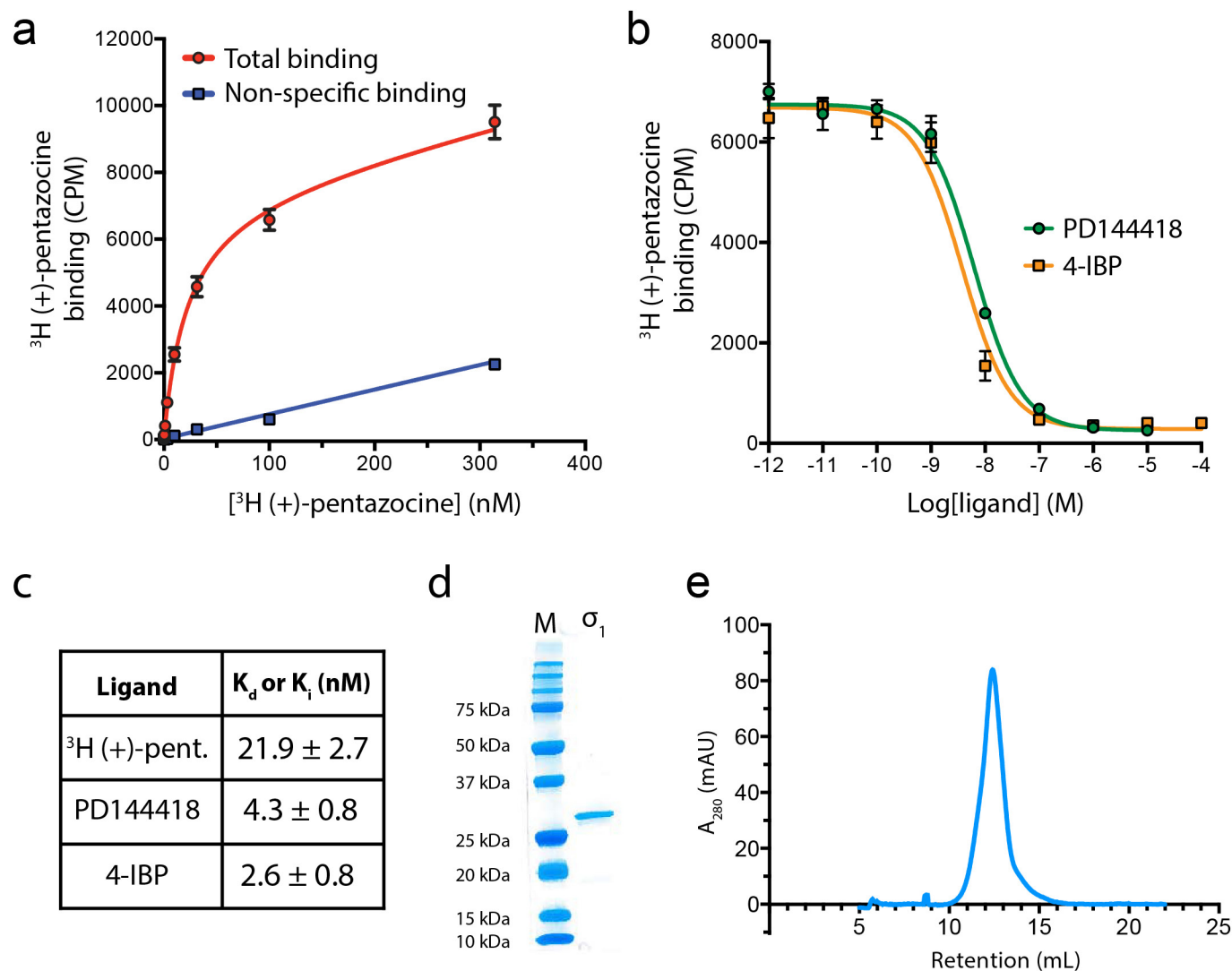
Oligomerization analysis. The oligomeric state of σ_1 receptor was assessed by SEC-MALS using a Wyatt Dawn Heleos II multi-angle light scattering detector and Optilab TrEX refractive index monitor with an Agilent isocratic HPLC system. Receptor was prepared as described above, but with no ligand added during purification. The ligand-free receptor was diluted to 0.5 mg ml $^{-1}$ in SEC-MALS buffer (0.025% *n*-dodecyl maltoside, 20 mM HEPES pH 7.5, 100 mM sodium chloride). Ligands were added to a final concentration of 25 μM to ensure stoichiometric excess over receptor and the sample was incubated with ligand at least 2 h at room temperature (21°C). Separation steps were performed in SEC-MALS buffer with a Tosoh G4SWxl column at a flow rate of 0.5 ml min $^{-1}$. Data analysis used the Astra software package version 6.1.4.25 (Wyatt) using the protein conjugate method with previously reported dn/dc values for detergent⁴². The effect of ligands in LMNG/CHS mixed micelle buffer was examined by analytical size exclusion in a similar procedure. The receptor was incubated with a twofold stoichiometric excess of the appropriate ligand and then subjected to SEC on a Superdex 200 column in a buffer consisting of 100 mM sodium chloride, 20 mM HEPES pH 7.5, 0.01% LMNG, 0.001% CHS, $1\ \mu\text{M}$ ligand.

Oligomerization state was also assessed by native PAGE. For these experiments, 7.5 μg of σ_1 receptor was mixed with tenfold stoichiometric excess of SKF10,047 or NE-100 in a 10 μl reaction containing 20 mM HEPES pH 7.5, 250 mM NaCl, 0.1% MNG, 0.001% CHS. After incubation at room temperature for 1 h, the reaction was added to 1 μl loading buffer consisting of 50% glycerol and 0.25% (w/v) bromophenol blue and separated by 10% native PAGE running in Tris-glycine (pH 8.3) buffer supplemented with 0.5% CHAPS and 0.5% sodium cholate for 4 h at 150 V in an ice bath. Blue native PAGE was performed as previously described⁴³. In brief, 10 μl reaction was supplemented with 1 μl of 50% glycerol and 1.5 μl of 0.1% Coomassie blue G-250 and was loaded onto a linear 3–12% gradient native PAGE gel (Life Technologies) running in blue cathode buffer supplemented with 0.05% MNG, 0.0005% CHS for 4 h at 150 V in an ice bath. The gel was stained using an InstantBlue staining kit (CBS Scientific).

Radioligand binding. Radioligand binding experiments were performed similarly to established procedures²³. In brief, *Sf9* membranes expressing σ_1 receptor were prepared by dounce homogenization followed by centrifugation. Resuspended membranes were aliquoted and flash frozen before use. For each binding experiment, membranes were incubated with ^3H (+)-pentazocine (Perkin Elmer) at the indicated concentration or at a fixed concentration of 10 nM for competition binding assays. To approximate physiological conditions, incubation was performed at 37°C for 2 h in 150 mM sodium chloride,

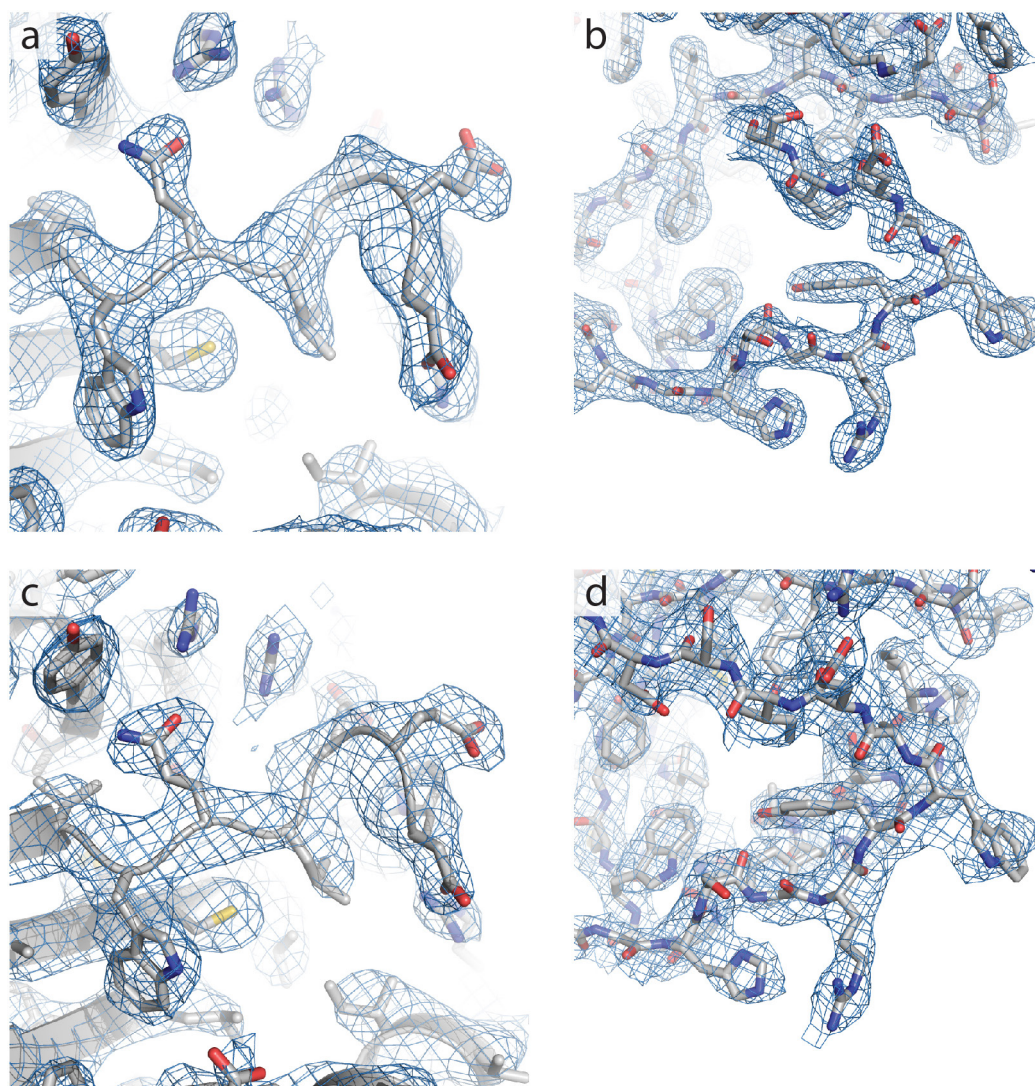
20 mM HEPES pH 7.5 and 0.1% (w/w) bovine serum albumin. Filter pads were incubated with 0.3% polyethyleneimine for 20 min, then samples were loaded onto the filter and washed using a Brandel harvester. Radioactivity was quantified by liquid scintillation counting. Non-specific binding was quantified by replicate reactions in the presence of 2 μ M haloperidol. All measurements were performed in triplicate and repeated in two independent experiments. Experiments in Tris pH 7.5 showed similar results to those conducted in HEPES. Data analysis used GraphPad Prism, with K_i values calculated by Cheng-Prusoff correction using the experimentally measured probe dissociation constant.

31. Caffrey, M. & Cherezov, V. Crystallizing membrane proteins using lipidic mesophases. *Nature Protocols* **4**, 706–731 (2009).
32. Rasmussen, S. G. *et al.* Structure of a nanobody-stabilized active state of the β_2 adrenoceptor. *Nature* **469**, 175–180 (2011).
33. Bricogne, G., Vonrhein, C., Flensburg, C., Schiltz, M. & Paciorek, W. Generation, representation and flow of phase information in structure determination: recent developments in and around SHARP 2.0. *Acta Crystallogr. D* **59**, 2023–2030 (2003).
34. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
35. Afonine, P. V. *et al.* Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D* **68**, 352–367 (2012).
36. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
37. Schrodinger, LLC. The PyMOL Molecular Graphics System, v.1.3r1 (2010).
38. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
39. Morin, A. *et al.* Collaboration gets the most out of software. *eLife* **2**, e01456 (2013).
40. Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* **38**, W529–W533 (2010).
41. Holm, L. & Rosenström, P. Dali server: conservation mapping in 3D. *Nucleic Acids Res.* **38**, W545–W549 (2010).
42. Strop, P. & Brunger, A. T. Refractive index-based determination of detergent concentration and its application to the study of membrane proteins. *Protein Sci.* **14**, 2207–2211 (2005).
43. Reisinger, V. & Eichacker, L. A. Analysis of membrane protein complexes by blue native PAGE. *Proteomics* **6** (Suppl. 2), 6–15 (2006).



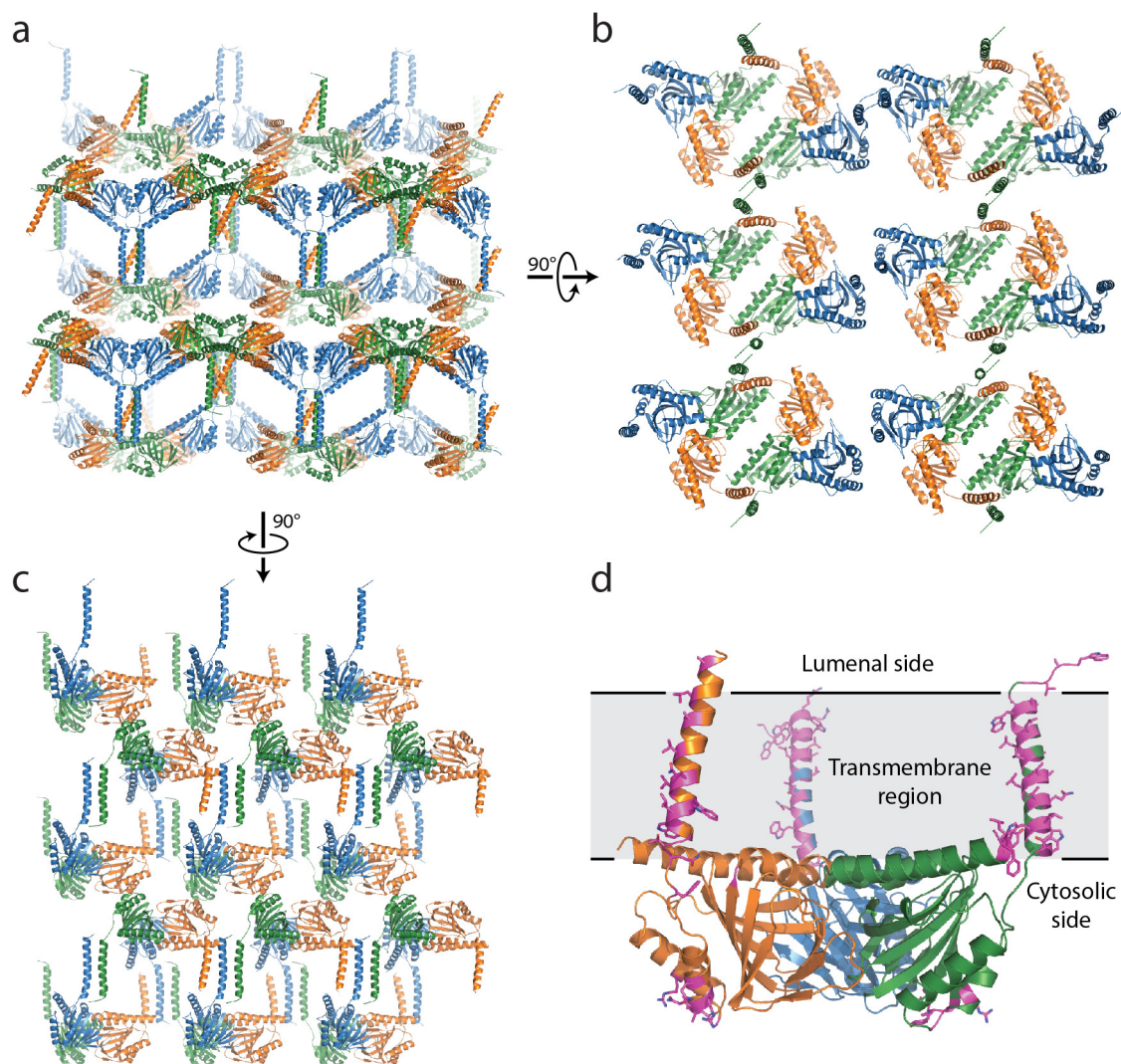
Extended Data Figure 1 | Assessment of σ_1 functional properties and biochemical quality. **a**, Saturation binding curve to measure K_d for ^3H (+)-pentazocine, with points shown as mean \pm s.e.m. **b**, Competition binding measurement of affinities for the two co-crystallized ligands with

points shown as mean \pm s.e.m. **c**, Summary of binding affinities with 95% confidence intervals for K_d/K_i values. **d**, Analysis of receptor purity by SDS-PAGE. **e**, Analytical size exclusion of purified σ_1 receptor in LMNG/CHS detergent buffer on a Superdex 200 column.



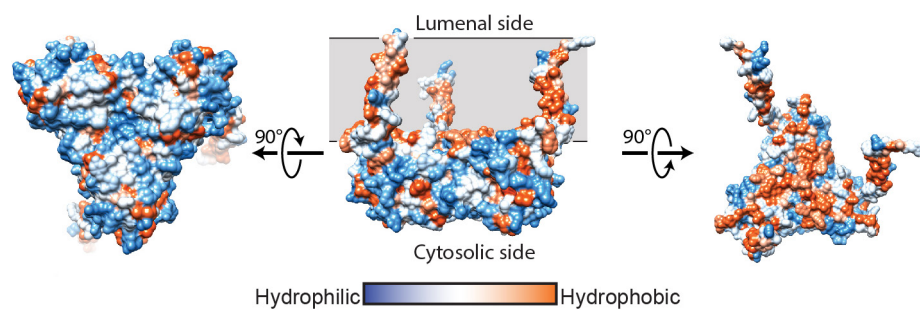
Extended Data Figure 2 | Representative electron density. **a**, Composite omit $2F_o - F_c$ electron density contoured at 1.0σ for σ_1 receptor bound to PD144418, showing a loop from Val73 to Glu78 as well as surrounding

residues. **b**, The same map over a loop from His116 to Ser125. **c**, The equivalent map to that in **a**, calculated for σ_1 receptor bound to 4-IBP. **d**, The equivalent map to that in **b**, calculated for σ_1 receptor bound to 4-IBP.

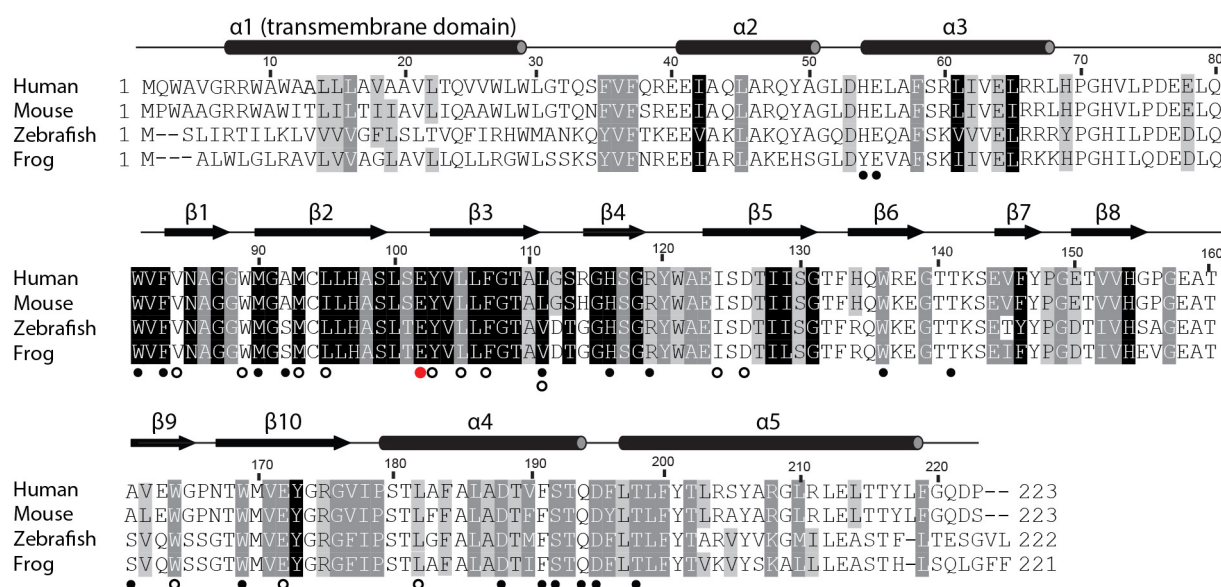


Extended Data Figure 3 | Lattice contacts. **a**, Lattice packing of the σ_1 receptor viewed parallel to the membrane plane. **b**, **c** A view normal to the membrane and another parallel view, respectively. **d**, A single σ_1 trimer is

shown, with lattice contact residues highlighted in magenta sticks. Lattice contacts are formed primarily through interactions of the relatively poorly conserved transmembrane helices.

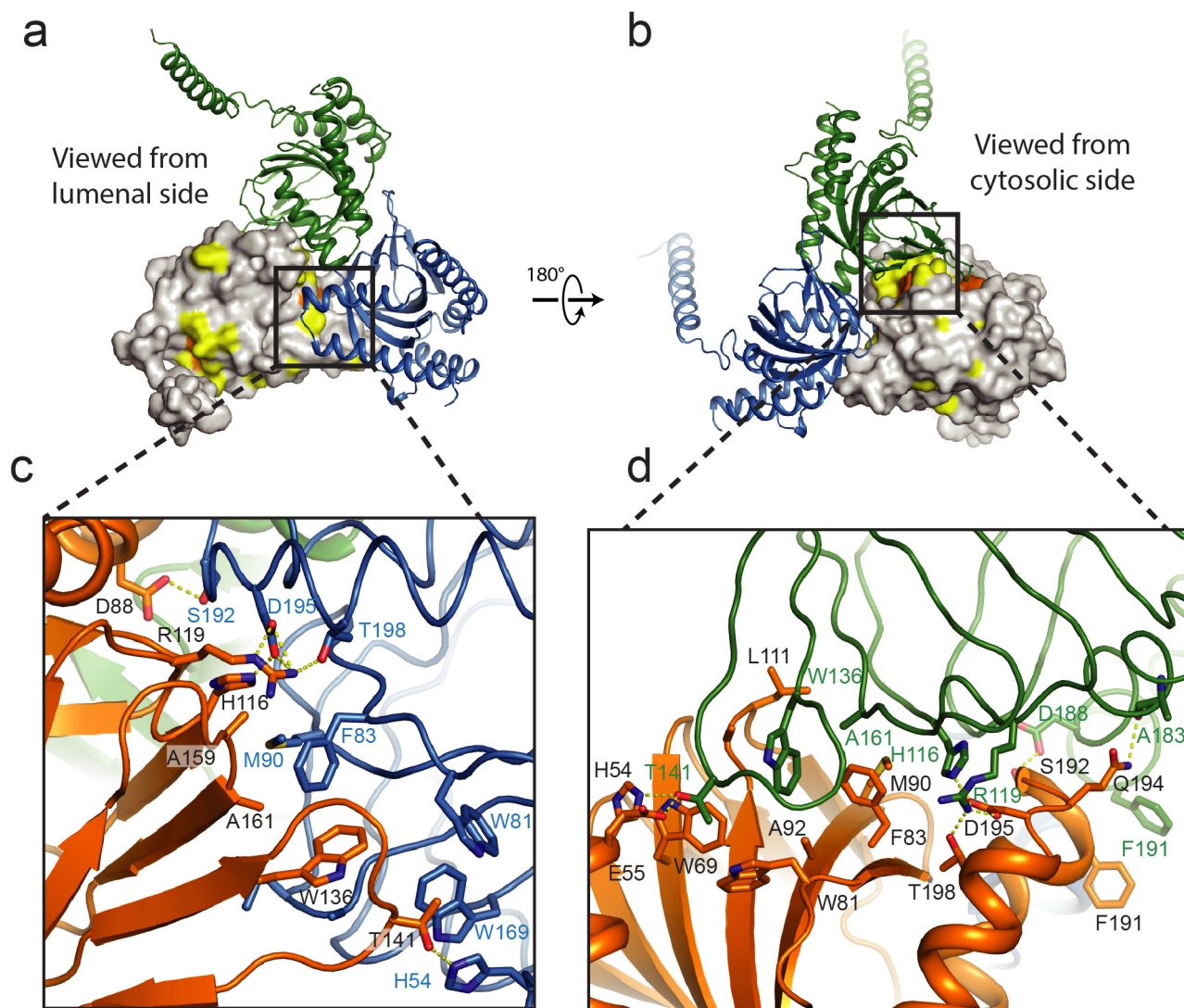


Extended Data Figure 4 | Hydrophobicity analysis. a. The structure of σ_1 receptor shows a hydrophilic (blue) surface on the cytosolic face (left), while transmembrane domains and the membrane-facing surface of the receptor trimer are hydrophobic (orange; right panels). Hydrophobicity analysis was conducted using UCSF Chimera.



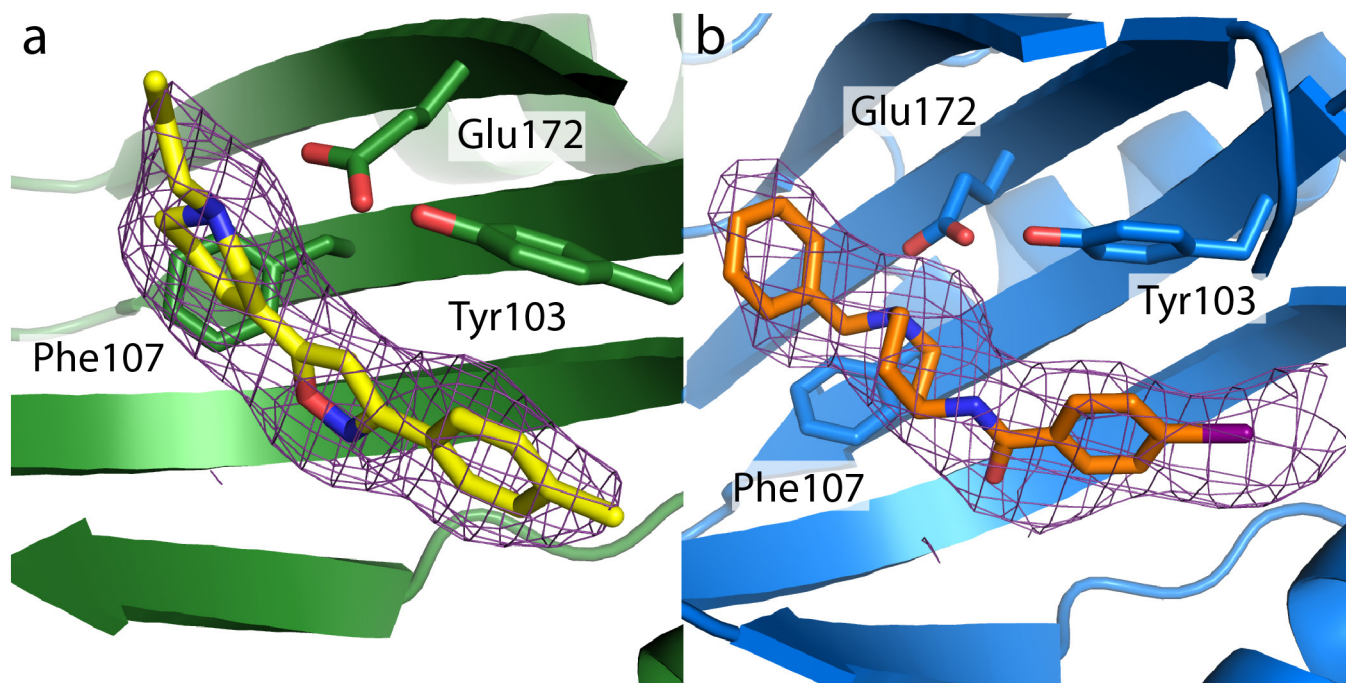
Extended Data Figure 5 | Sequence conservation. The results of an alignment of 277 sigma receptor sequences from vertebrates with *Homo sapiens*, *Mus musculus*, *Danio rerio*, and *Xenopus laevis* displayed. Residues with 98%, 80%, and 60% similarity are shown in black, grey, and light grey respectively. Secondary structure elements are shown above the alignment

on the basis of the human σ_1 receptor crystal structure. Open black circles mark residues within 4 Å of the ligand-binding site, solid black circles below the alignment denote residues located in the trimerization interface, and a red circle marks the site of the E102Q mutation associated with amyotrophic lateral sclerosis.

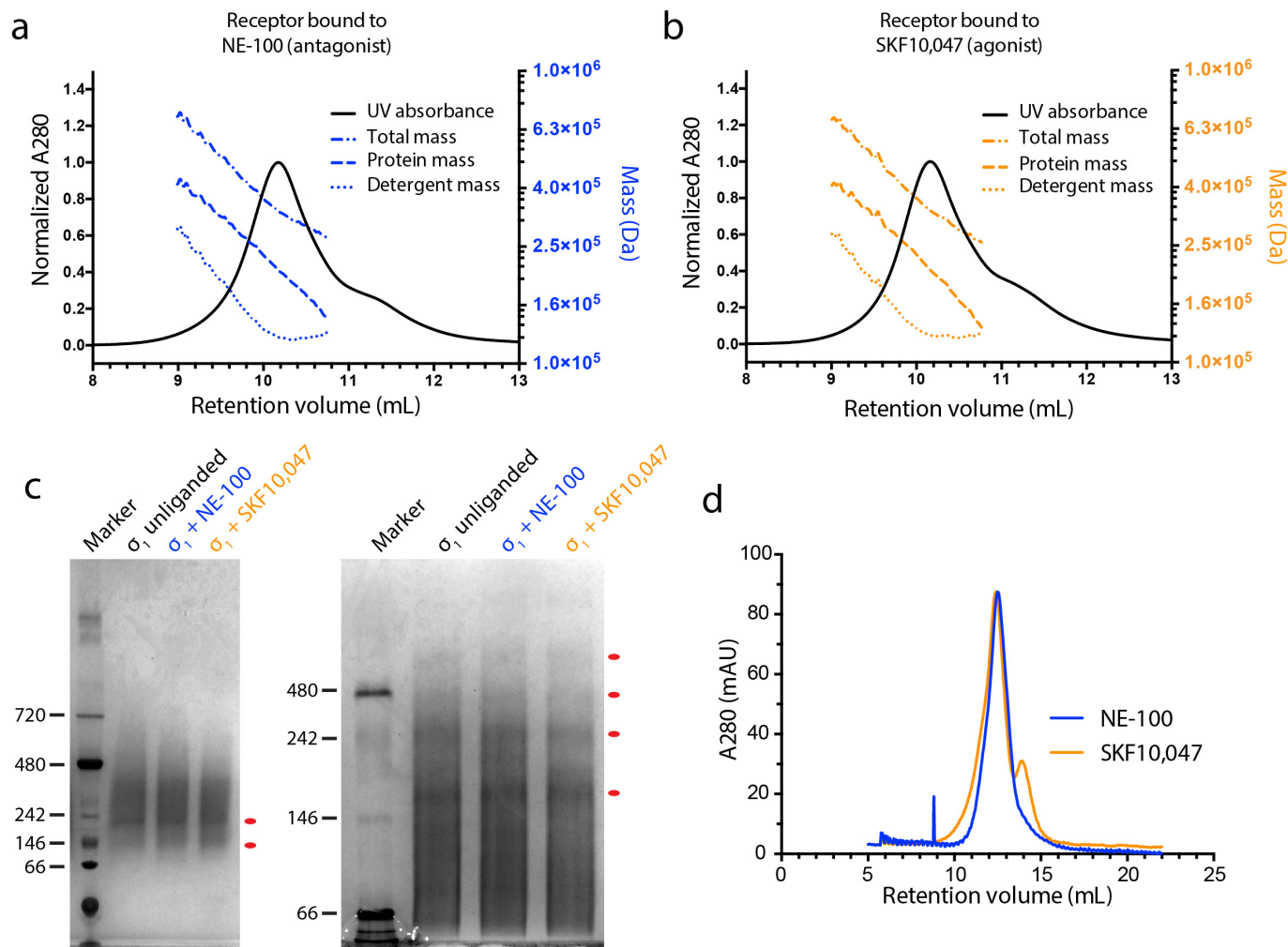


Extended Data Figure 6 | Trimerization interface. a, b, Two views of the trimerization interface are shown, coloured by sequence conservation. Residues highlighted in yellow are more than 80% conserved among a selection of 300 σ_1 receptor homologues, and residues in orange

surface are more than 98% conserved. c, d, Close-up views of the interface, showing the extensive hydrophobic and polar contacts at the oligomerization interface.



Extended Data Figure 7 | Omit maps of PD144418 and 4-IBP. **a**, An $F_o - F_c$ omit map contoured at 1σ showing the electron density (purple) of PD144418 (yellow). **b**, An equivalent map showing the electron density (purple) of 4-IBP (orange).



Extended Data Figure 8 | Oligomerization state. **a**, Analysis of receptor oligomerization by SEC–MALS in the presence of the classical antagonist NE-100 or **(b)** the classical agonist SKF-10,047. The peak is 38% detergent and 62% protein by mass. The total mass of each component varies throughout the peak, indicating a mix of oligomeric species. **c**, Analysis of oligomerization state by blue native PAGE (left) and a higher-resolution

detergent-supplemented tris-glycine native PAGE gel (right), showing a similar polydisperse profile. Discrete oligomers are marked with red dots, corresponding to possible trimers, hexamers, and higher-order species. **d**, In a mixed micelle of lauryl maltose neopentyl glycol and cholesterol hemisuccinate, modest differences in SEC profile are observed between agonist- and antagonist-treated receptor.

Extended Data Table 1 | Data collection and refinement statistics

	σ_1 bound to PD144418 (native)	σ_1 bound to 4-IBP	σ_1 bound to PD144418 Ta ₆ Br ₁₂ soak
Data collection^a			
Wavelength (Å)	1.033	1.033	1.2548
Space group	P 2 ₁ 2 ₁ 2	P 2 ₁ 2 ₁ 2	P 2 ₁ 2 ₁ 2
Number of crystals	1	3	1
Unit cell dimensions			
<i>a</i> , <i>b</i> , <i>c</i> (Å)	85.6, 126.1, 109.7	85.7, 126.8, 110.8	85.0, 127.4, 109.4
α , β , γ (°)	90, 90, 90	90, 90, 90	90, 90, 90
Resolution (Å)	40 - 2.5 (2.65 - 2.50)	50 - 3.2 (3.30 - 3.20)	46.2-3.50 (3.63-3.50)
Completeness (%)	98.8 (97.7)	97.1 (97.8)	99.0 (99.2)
<I/ σ (I)>	10.1 (0.9)	5.8 (1.9)	12.4 (3.1)
CC _{1/2} (%)	99.8 (40.1)	98.0 (35.6)	98.2 (12.9)
Multiplicity	3.5 (3.4)	3.2 (3.2)	3.1 (3.1)
Refinement			
Resolution (Å)	40 - 2.51 (2.57 - 2.51)	33.6 - 3.2 (3.28 - 3.20)	
No. reflections	41026 (2000 in test set)	19968 (1998 in test set)	
R _{work} /R _{free} (%)	19.5/23.3	21.8/26.1	
No. atoms			
Protein	5097	5027	
Ligand	63	69	
Solvent ions/lipid	160	130	
Water	136	0	
B factors (Å ²)			
Protein	79.5	66.5	
Ligands	84.6	90.2	
Water	76.8	N/A	
Solvent ions/lipids	116.8	96.6	
RMS deviation			
Bond length (Å)	0.003	0.003	
Bond angles (°)	0.586	0.613	
Ramachandran statistics ^b			
Favored	98.8%	98.4%	
Allowed	1.2%	1.6%	
Outliers	0%	0%	

Extended Data Table 2 | Structural homologues of the σ_1 receptor

PDB ID	Z-score	RMSD (Å)	Seq. ID to σ_1 (%)	Name and bound metal ion	Oligomerization state
3BCW	10.8	2.8	8	Unknown function cupin (no metal ion)	Dimer
2PFW	9.9	2.5	11	Unknown function cupin (no metal ion)	Dimer
4AXO	9.4	3.5	9	CD1908, a bacterial microcompartment for the breakdown of ethanolamine (no metal ion)	Hexamer
4BIF	9.4	2.6	12	Manganese- dependent hydroxynitrile lyase (Mn)	Tetramer
1VJ2	9.3	2.7	10	Manganese-containing cupin (Mn)	Dimer
4UXA	9.3	2.7	12	(R)-selective manganese-dependent hydroxynitrile lyase (Mn)	Dimer
2Y0O	9.2	2.6	10	Probable D-lyxose ketol isomerase (Zn)	Dimer
4E2G	9.2	2.9	8	Cupin fold protein Sthe2323 (Ni)	Dimer
2OYZ	9.2	2.5	11	Unknown function protein VPA0057 (no metal ion)	Dimer
4QM8	9.1	2.7	6	Cysteine dioxygenase (Fe)	Monomer
3LWC	9.1	2.5	11	Unknown function (no metal ion)	Dimer
3EBR	9.0	2.9	12	Rmlc-like cupin protein (no metal ion)	Tetramer
1O4T	9.0	2.6	15	Predicted oxalate decarboxylase (Mn)	Dimer
2F4P	9.0	2.4	11	Hypothetical protein TM1010 (no metal ion)	Dimer
5BPX	8.9	2.9	15	2,4'-dihydroxyacetophenone dioxygenase (Fe)	Dimer
3HT2	8.9	3.3	11	Zinc containing polyketide cyclase RemF (Zn)	Dimer
2OPK	8.9	2.2	14	Putative mannose-6-phosphate isomerase (no metal ion)	Dimer
3BAL	8.8	2.9	15	Acetylacetone dioxygenase (Zn)	Tetramer
1YLL	8.8	3.4	6	Unknown function PA5104 (no metal ion)	Tetramer
4E2S	8.7	4.2	9	(S)-ureidoglycine aminohydrolase (Mn)	Octamer
3ESG	8.7	3.1	8	HutD (no metal ion)	Dimer
3L2H	8.7	3.1	14	Putative sugar phosphate isomerase (no metal ion)	Tetramer
1H1I	8.6	3.4	3	Quercetin 2,3-dioxygenase (Cu)	Dimer
1GQG	8.4	3.4	3	Cu-dependent Quercetin 2,3-Dioxygenase (Cu)	Dimer
3SCH	8.4	3.0	9	Hydroxypropylphosphonic acid epoxidase (Fe)	Tetramer
3KMH	8.3	3.0	7	Sugar isomerase (Mn)	Dimer
1V70	8.2	3.0	9	Probable antibiotics synthesis protein (Na)	Dimer
4LA2	8.2	2.9	15	Dimethylsulphoniopropionate lyase (Zn)	Monomer

Structural homologues of the σ_1 receptor were identified by search with the DALI server, and those with Z-score values above 8 are summarized here. All are cupin fold proteins, with most showing oligomeric structures based on annotated biological assembly in the Protein Data Bank. Trimeric structures such as that seen for the σ_1 receptor have not been reported previously for other cupin-fold proteins.

CAREERS

OCCUPIED The ‘carp lady’ at Malheur National Wildlife Refuge debriefs **p.533**

JOB SKILLS How to answer the classic interview questions go.nature.com/h4nabz

NATUREJOBS For the latest career listings and advice www.naturejobs.com

MATT WYCZALKOWSKI



Members of the Balsa (Biotechnology and Life Sciences Advising) consulting group in St Louis, Missouri. The group mostly consists of PhD students.

NON-PROFIT WORK

Take my advice

Part-time work at a consulting firm can provide management skills and connections for graduate students and postdocs who hope to move beyond academia — or not.

BY CHRIS WOOLSTON

Not many graduate students who spend 50–60 hours in the laboratory each week are eager to take on an outside job — especially one that pays nothing. But Michael Lang, a PhD student in cell and developmental biology at the University of Michigan in Ann Arbor, has added two part-time, unpaid positions to his workload. He's the president and co-founder of miLEAD Consulting, an independent, non-profit company based in Ann Arbor that connects the university's graduate students and postdoctoral researchers with local biotechnology and health-care companies that need help with product development, market analysis or branding. And he

works directly for miLEAD to provide his own insights and analyses to companies.

Lang thinks that the long hours are worth it. The consulting work helps him to build leadership and management skills that would come in handy if he were to reach his ideal goal of running an academic lab. And if that doesn't work out, he'll have a fall-back position: "I've always wanted to be a scientist, but a US\$130,000 job at a top consulting firm sounds pretty good too."

Lang's group is one of several consulting organizations that have sprung up on US campuses in the past few years. They supply teams of postdocs and graduate students who can take a scientific approach to common questions faced by local biotechnology and pharmaceutical

start-ups — what is the demand for a new product, what is the competition, what can be done to make a product better and what is the best way to profit from a good idea? Consultants do not always know how companies use their input or whether their advice makes a difference, but the value of the experience is undeniable. "We want to give people another bullet point on their CV," Lang says. "It can get them over the hurdle to getting a job."

A few of these consulting groups, including miLEAD, are independent, non-profit companies with no official ties to their home institute. But most are affiliated with their host institutions, including Harvard University in Cambridge, Massachusetts, Stanford University in California and the University

► of Pennsylvania in Philadelphia. Such campus-based organizations haven't caught on outside the United States, but at least one global company, 180 Degrees Consulting, recruits postdocs and graduate students for consulting projects and gives scientific trainees in the United Kingdom and elsewhere a chance to add to their skill set.

Whatever group they work for, trainees in consulting get valuable experience in analysis, decision making and team-based problem solving that can give them a boost in the job market. It is also a break from the normal routine. "Fast-paced teamwork can be a lot of fun," says Huadi Zhang, a medical-science PhD student and co-president of Harvard Graduate Consulting Club. "I didn't have that kind of experience in the laboratory." But on-the-side consulting is also a serious commitment and time drain — and there are several hoops to be jumped through if students want to start a group from scratch (see 'How to start a consultancy'). The field is not for everyone, but an increasing number of trainees have found that it is possible to consult their way into a career.

CV BOOSTER

For Lang, consulting has turned into a second life outside the lab. He estimates that he spends 10–15 hours a week fulfilling his duties as president of miLEAD: overseeing the search for clients, recruiting consultants and, importantly, training them in the basics of business. Working on a project — which might involve meeting with a company's board, talking to doctors or digging through research articles — generally takes him another 10–15 hours each week. These are huge time commitments for a graduate student with experiments to run and papers to write. But it's worth it, he says, for the boost it gives to his CV and research. "The additional work has helped me streamline my science," he says. "There's not a lot of downtime in the lab."

Lang's recent projects include an eight-week gig for a Michigan pharmaceutical company that is developing a therapeutic drug for newborns. (Because of non-disclosure agreements, he cannot name the company.) He and his team studied the market for the drug, scoped out the competition and gauged its potential applications in neonatal medicine. Previously, he was on a team that spent four weeks assessing an app-based learning tool for college students that was developed at the University of Michigan.

Lang says that miLEAD brought in \$6,000 in revenue in 2015 and is aiming for \$12,000 in 2016. The board uses all of the revenue for group-related activities, including flying in speakers for panel discussions and funding team-building gatherings. If the coffers get sufficiently full, Lang hopes to start a grant

"We treat this like a business. If money is involved, better work gets done."

Early-career researchers at institutions that do not have a consulting organization can start one themselves. The first step is evaluating and comparing existing groups to find a model that fits. Michael Lang, president of miLEAD Consulting in Ann Arbor, Michigan, recommends setting up a non-profit corporation that charges at least a nominal fee for its services.

Simran Madan, PhD student at Baylor College of Medicine and senior vice president of the Texas Medical Center's consulting group in Houston, says that it's important to survey the local scene to determine whether there are enough trainees around who have the time and interest for consulting work, and enough local businesses that could use help. She also recommends finding a confidante who has been through the process. "Since

programme to help local businesses to get off the ground. miLEAD's fees for client companies are a tiny fraction of what a big-time consulting company would charge, but they underscore the professionalism of the process. "We treat this like a business," he says. "If money is involved, better work gets done."

Conversely, Zhang says that the Harvard Graduate Consulting Club has no plans to start charging clients. "It's a way for us to give back to the community," he notes. Although it is likely that local start-ups get some value from their consulting, improving a company's bottom line is not the main point of the exercise. "It's a learning experience for us," says Zhang.

A GROWING FIELD

Consulting organizations are starting to pop up on other campuses, giving more postdocs and graduate students a chance to try out the field. Simran Madan, a PhD student in translational biology at Baylor College of Medicine in Houston, Texas, is helping to kick-start consulting services at the Consulting Club as its senior vice president at the Texas Medical Center in Houston. This independent, non-profit group is drawing talent from several local institutions, including Baylor and the University of Texas Health Science Center and MD Anderson Cancer Center in Houston. The group aims to begin offering consulting services by the end of the year. For now, Madan and club president Redwan Huq, a Baylor PhD student in molecular physiology and biophysics, are learning how to recruit potential consultants, provide training, structure consulting teams and attract clients.

The plan is to charge local companies about \$500 for 6 weeks of work analysing a product

CASE STUDY

How to start a consultancy

setting up a non-profit is a monumental challenge, we recommend consulting with someone who has the expertise," she says.

Madan suggests working with university administration to get their support; even though the non-profit group won't technically be a part of the campus, the approval and cooperation of an institution can be crucial for long-term success. It's also important to set up a team with sharply defined roles and a chain of command.

Paperwork is involved, not surprisingly. In the United States, it takes a lengthy and complicated application to the US Internal Revenue Service to obtain non-profit status. Among other things, the application must show that the organization will not make money for the founder. But once obtained, the status allows the group to accept donations and avoid paying income tax. **C.W.**

and coming up with a marketing or development plan, a price that should be attractive to cash-strapped start-ups. "Professional consultants are expensive, and you almost never see a start-up hiring a firm," Madan says. "But they can get the same sort of analysis from a trainee."

One source of inspiration for Madan and Huq is the BALSA (Biotechnology and Life Sciences Advising) group, a successful consulting organization at Washington University in St Louis, Missouri. BALSA, which started in 2011, has 100 active members who participate in around 40 projects a year. About 60% of the members are science PhD students, 30% are science postdocs and a few are business or law students. Each job lasts six weeks, and each team includes three consultants, a project manager and an adviser. Most of the work involves product development and market analysis for local start-ups and entrepreneurs in the biotechnology, agriculture and health-care industries. The group also has clients in South Dakota; San Francisco, California; and Philadelphia, Pennsylvania, says Shivam Shah, who is the BALSA president and a PhD student in biomedical engineering at Washington University.

A frequent BALSA client is Washington University's Office of Technology Management, which has often hired the team to help evaluate patent applications from faculty members. Shah says that the group tries to avoid having students evaluate their direct supervisors, but that is not always possible. Students aim to judge patent applications strictly on their scientific merit and real-world potential, he says.

Since joining the group in 2013, Shah has worked on more than 20 projects as either a consultant or a project manager. Working on multiple projects has given him a chance to

fine-tune his management style and learn more about the scientific marketplace, he says. He hopes to land a consulting job soon after getting his degree, perhaps with a health-care venture-capital firm looking for advice about wise places to invest.

But a consulting career is hardly the only destination for Balsa members. Many have ended up working in industry as research scientists, patent specialists or consultants for companies such as the multinational agrochemical company Monsanto, based in St Louis, Missouri, and the New York-based computing giant IBM. And of the roughly 200 alumni of the programme, he estimates that about one-third have continued in academic careers. The skills learned in the consulting game — management, leadership and teamwork — would prove valuable to anyone running their own lab, Shah says.

There is a paucity of organizations such as miLEAD and Balsa outside the United States, but early-career scientists in the United Kingdom, Europe and elsewhere can still get real-life consulting training. One option is a position with 180 Degrees Consulting, a global organization with branches in Cambridge, UK; King's College London; Munich, Germany; the University of Tokyo; the University of Sydney; and the University of California, Los Angeles, among many other sites. The company enlists students and postdocs to provide pro bono consulting to non-profit and humanitarian organizations around the world. Although the work generally is not focused on scientific issues, science PhD students and postdocs can bring valuable skills to the organization, says Daniel Jiang, a PhD student in computer science who in 2015 founded the 180 Degrees Consulting branch at King's College London. "I know more about data sets than a political-science major does," he says.

Jiang's group is working with a children's charity and sports charity in London, and a school in the Philippines. The company attracts people who want to make a positive difference in the world, Jiang says, but there are benefits for the consultants themselves. "It's a great opportunity for students to find out about a different career before they graduate," he says.

Lang of miLEAD is still technically a student, but he's racking up professional-grade experience and isn't slowing down: he'll jump into two new projects as an adviser this summer. He can't discuss details, but the big picture is clear: he'll be working long hours, thinking about tough problems and moving closer to a postgraduate career.

Are the long days worth it? That's a cost-benefit analysis that he has figured out on his own, no consultant required. ■

Chris Woolston is a freelance writer in Billings, Montana.

TURNING POINT

Carpe freedom



Armed militants who were protesting against how public land is managed took over Oregon's Malheur National Wildlife Refuge on 2 January. They stayed for 41 days and caused roughly US\$6 million in damages. US Fish and Wildlife Service biologist Linda Sue Beck describes the occupation and its aftermath.

Did you anticipate the takeover?

No. We knew that the militia was in town for a peaceful march to protest against the prison sentence of a father and son convicted for arson on federal lands, but we didn't expect anything like what happened. The occupiers did a lot of damage, to our offices and the land, as well as to tribal archaeological artefacts.

What management issues do you work on?

The refuge was established in 1908 to support millions of resident and migratory birds. In the 1920s, someone brought common carp (*Cyprinus carpio*) into the basin, and they've become a problem. Before they were introduced, 9 species of submerged aquatic vegetation covered 90% of the lake. That, and the associated macroinvertebrates, drew birds. Today, the common carp are in direct competition with the birds for that food. They also muddy the water so that there is no light for the plants to grow. We're trying our best to get the carp under control.

Is federal-land management contentious?

There have been contentious issues, but my experience has been mostly positive. Together with tribal members, ranchers, non-governmental organizations and other government agencies, we spent 5 years over 40 meetings to write a 15-year plan for the refuge. People were vocal about things they didn't like, but in the end, the number-one priority was carp control. We agreed that we want it to be healthy again so that it can serve as a grocery store for the birds.

Your name appeared in news reports, as if the militants were targeting you. Why?

Essentially, they were sitting at my desk. At one point, a news article suggested that I was one of the reasons the occupation was happening. I've never had a rancher call me out — I have no idea where that came from. And to be honest, it freaked me out when my parents were contacted by a journalist. Then another person wrote an article entitled, 'I stand with Linda Sue Beck'. I think I was just the target for news that day.

That piece gained traction on social media.

What was it like?

It was nice to have support. I also have a good relationship with the locals, in part because I've involved them in science experiments where, for example, the public catches fish so that I can collect data. Some local ranchers turn carp into an organic fertilizer to use on their fields. The militants picked the wrong refuge to take over. I think they thought it would be easier to sway the locals, but our partnerships are strong. People are sending cheques from all over the world. Hopefully, we can use those funds to get the refuge back up to what it was.

What was the first day back at work like?

We had to evacuate the area after the takeover, and I was sent to our office in Vancouver, Washington, until the occupation was over. Coming back for the first time, I had to go through two FBI roadblocks and be escorted to my heavily guarded office. We're still piecing together the full impact of the damage.

How did the takeover affect your work?

We missed an opportunity to remove thousands of carp from the lake. In December, the lake was at a record low of about 800 hectares, so we had planned to block carp while they were aggregated at the mouth of the river, so that we could pull them out of the system. The lake has since grown to roughly 8,000 hectares, and the fish have dispersed because it is so deep.

How did it affect your outlook?

I realized how important it is to be honest and to keep lines of communication open. My approach to science is that I believe in what I'm doing to conserve land and animals for future generations. There might be political stuff at play, but I do what is best for the birds. ■

INTERVIEW BY VIRGINIA GEWIN

This interview has been edited for length and clarity.

A SLICE OF TIME

You must remember this ...

BY JEFF HECHT

The Oligarch sitting behind the vast and finely polished desk of genuine red mahogany looked very small or very far away or both. It had taken me a long time to get this far, and I knew that I needed to wait for her to speak.

"I understand that you have something interesting to offer me," she said. My eyes could not focus sharply on her face, but I could see her lips move. Her voice sounded so perfect it had to be synthesized. My ears could not tell from what direction the words really came, but I had expected something like that.

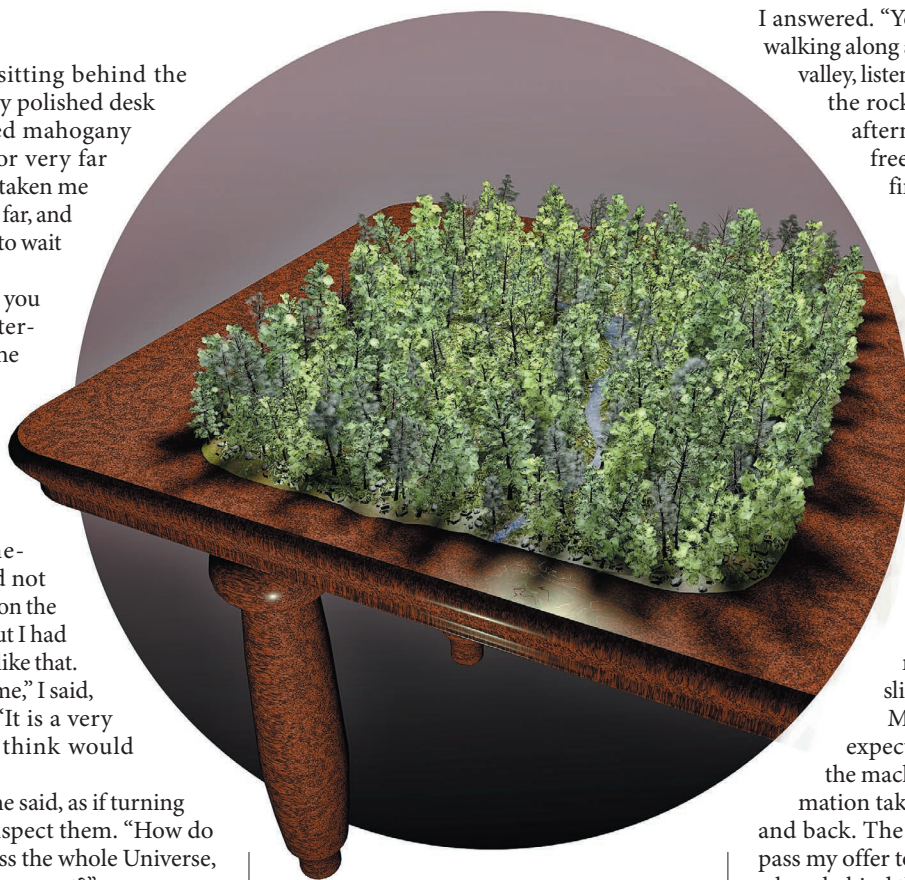
"I have a slice of time," I said, pausing for effect. "It is a very special slice that I think would interest you."

"A slice of time," she said, as if turning the words over to inspect them. "How do you slice time? Across the whole Universe, or across some little segment?"

Perceptive, I thought. And probing. "Only across the portion of the Universe that you affect," I answered. As an Oligarch, she affected all that humanity could reach, but our part of the Universe was very small. Like the rest of us, she had no way to the stars. Our ancestors had studied them for untold millennia, and had even seen their planets. Yet in the Age of the Oligarchs we know our limits. The stars are but bright spots in the sky that we can see but never touch.

Her eyes probed me. I wondered where she really was and how many layers of simulation there were between us. She was not an original Oligarch; they had lived and died before the later Oligarchs had learned how to extend their lives almost indefinitely. Their minds were copied into simulations, and their bodies put into near-immortal suspension, to be disturbed only when the simulations could not answer. Otherwise, their subordinates ruled, consulting with the simulations, until it came time for the subordinates themselves to be suspended, and for their subordinates to rule.

I had spent months showing my slice of



time to layers of junior subordinates and under-assistants before I could reach her simulation. I waited for her words to come.

"We have no record of any missing time," she said at last, the faint wisps of vibrating air molecules spreading through the room, then sinking into the walls without an echo.

"This slice comes from before the records of time began," I said, stopping so my words could spread through the room in the same way, and so the wheels within wheels of her simulated thoughts could spread through the network of machines that are her interface with the world. I wanted them to reach back to whatever human part of her existed hidden somewhere deep and dark and safely sequestered.

Time passed, as I knew it would. Thought takes time, and human thought takes more time than simulations that can ponder only a finite library of possibilities.

"When?" she asked at last.

"You were a child,"

I answered. "Young, perhaps four or five, walking along a stream in a little mountain valley, listening to the water tumble over the rocks on a sunny late summer afternoon. You were happy and free. It was not an easy time to find."

"Show me," she said.

I opened the slice of time, spreading it wide on the vastness of the table. It unfolded larger and larger, taking on more reality as it grew. The water and the rocks sparkled in the sunlight that came through the leaves; the air was warm and fresh. I could feel a light breeze and hear the water tumbling over the rocks in the stream. It had taken me many years to capture that slice of time.

More time passed, as I had expected. Latency, the masters of the machines call it. The time information takes to get from here to there and back. The time for the simulation to pass my offer to the ancient human somewhere behind the layers of simulations and human subordinates devoted to preserving their own lucrative servitude, and the time for her to send her response back to the simulation. They might delay but they had to obey; the ultimate control was hers. Was supposed to be hers, I told myself, hoping that she was not just dust inside an empty shell somewhere deep inside the vastness of the Oligarchy.

"You may have your price," came her voice, suddenly younger and full of hope.

I smiled. "You will enjoy it," I said, hoping that neither of us was lying. For my price was to be placed high in the line of her subordinates, just below the lowest level that had been frozen into the ranks of power and simulation. There I would hold the power of trusteeship in her name and in the names of all the higher subordinates in the frozen hierarchy of power, until the fullness of age came upon me and I sought my own slice of time. ■

Jeff Hecht is Boston correspondent for *Nature Scientist* and a contributing editor to *Laser Focus World*.

ILLUSTRATION BY JACEY

➔ NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm

nature INDEX 2016 SAUDI ARABIA

Nurturing home-grown
talent to solve local
problems

Forging global
connections sets the
pace of change

The institutes leading
the rise in output


DRAWING ON NEW RESERVES

*Oil wealth fuels science
ambitions*



Produced with
support from



nature publishing group 

nature INDEX 2016 SAUDI ARABIA

NATURE, VOL. 532, ISSUE NO. 7600 (APRIL 28, 2016)

Saudi Arabia is the world's largest oil producer, however, in 2008 the country announced its plans to broaden its scope from an oil-based economy to one based on knowledge, under a national science strategy laid out until 2030. Enormous investments have allowed the country to create new high-tech universities and establish cutting-edge laboratories in its leading research institutions.

Tracking the change in the country's scientific output over the past four years, the first Nature Index supplement about the Middle East shows that Saudi Arabia's investments in science are starting to pay off. It has surpassed all other Arab states in the region, and outstripped other regional leaders to achieve the second highest output in the index in Western Asia.

Most of this growth can be attributed to five institutions across the Kingdom, which we profile on page S8. These institutions have forged collaborations with 89 countries in 2015, and the research has fuelled the country's rapid rise in the index. The United States and China, the global leaders in science, continue to account for the bulk of collaborations with Saudi Arabia.

Page S16 describes how the country's leading institutions, KAUST and KAU, are driving these collaborations through two distinctly different approaches. Our feature on page S19 shows how one Saudi Arabian institution is reaping the great rewards of maintaining a focus on the

domestic and regional front.

A closer look at the Nature Index shows that, with the exception of KAUST, Saudi institutions usually have a smaller contribution to collaborations compared to their international counterparts. Aware of the need for more homegrown talent, the Saudi government has created a large scholarship programme to send students abroad for postgraduate studies, in the hope they will return and lead the country's scientific endeavours through new connections and a broad experience.

With the fastest growing weighted fractional count (WFC) in the Middle East in 2015, Saudi Arabia has positioned itself as a regional leader in the index. This rise has been driven by a strong focus on chemistry research, which makes up two-thirds of the Kingdom's science output in the Nature Index. On page S13, we look at the rise in Saudi Arabia's international standing. It has already shot past most of its competitors from 2012, and now has its sights on the higher rankings of the leading Asian strongholds. It's a lofty target, but the country's impressive trajectory so far makes it one to watch.

For more information on how Nature Index metrics are calculated, see page S24.

We acknowledge the financial support of KACST in producing this Nature Index supplement. *Nature* retains sole responsibility for all content.

Mohammed Yahia
Chief Editor, Nature Middle East

CONTENTS

- S2 SIGHTS SET ON A CENTRAL ROLE**
A graphic view of Saudi Arabia's growth in index output and increased connections
- S4 A 21ST CENTURY TRANSFORMATION**
Changing priorities at the centre of a quest to develop a knowledge economy
- S10 MAKING THE MOST OF FINANCIAL MIGHT**
Driven by a chemistry focus, Saudi Arabia sets new standards for the region
- S13 OILING THE WHEELS ON A ROAD TO SUCCESS**
A sustainable plan and significant funding offers strong international advantage
- S16 SHARED KNOWLEDGE IS KEY TO A KINGDOM**
Impressive results from global partnerships growing in scope and scale
- S19 MAKING THE MOST OF LOCAL EXPERTISE**
Joint attempts to solve regional problems often form basis for important breakthroughs
- S22 THE TABLES**
Institution rankings in the Nature Index by output and by subject performance
- S24 A GUIDE TO THE NATURE INDEX**
How to get the most out of the index and an explanation of the metrics

COVER IMAGE

Riyadh's skyline with the Kingdom and Al Faisaliyah centres, the country's third and fourth tallest buildings.



AJAL MUBARAK/GETTY IMAGES

EDITORIAL: Stephen Pincock, Mohammed Yahia, Sedeer El-Showk, Nadia El-Awady, Pakinam Amer, Rebecca Dargie, Victoria Kitchener. **ANALYSIS:** Larissa Kogleck. **ART & DESIGN:** Alisdair Macdonald, Kate Duncan. **WEB & DATA:** Bob Edenbach, Olivier Lechevalier, Naomi Nakahara, Pamela Sia, Bart Riepe, Jörn Ishikawa, Yuxin Wang, Jyoti Miglani, Jennie Pao, Paul Glaeser, Akiko Murakami, Takeshi Ouchi. **PRODUCTION:** Sue Gray, Karl Smart, Ian Pope, Matt Carey, Manpreet Mankoo. **MARKETING:** Adil Jouhadi, Alan Abery. **PROJECT MANAGER:** Anastasia Panoutsou. **SALES:** Jon Giuliani. **ART DIRECTOR:** Kelly Buckheit Krause. **PUBLISHING:** Nick Campbell, Richard Hughes, David Swinbanks.

NATURE INDEX 2016 SAUDI ARABIA

The Nature Index 2016 Saudi Arabia, a supplement to *Nature*, is produced by Nature Publishing Group, a division of Macmillan Publishers Ltd. This publication is based on data from the Nature Index, a website maintained by Nature Publishing Group and made freely available at natureindex.com.

Nature Editorial Offices
The Macmillan Building
4 Crinan Street,
London N1 9XW, UK
Tel: +44 (0)20 7833 4000
Fax: +44 (0)20 7843 4596/7

CUSTOMER SERVICES

To advertise with the Nature Index, please visit natureindex.com/client-services
feedback@nature.com
Copyright © 2016 Nature Publishing Group.
All rights reserved.

SIGHTS SET ON A CENTRAL ROLE

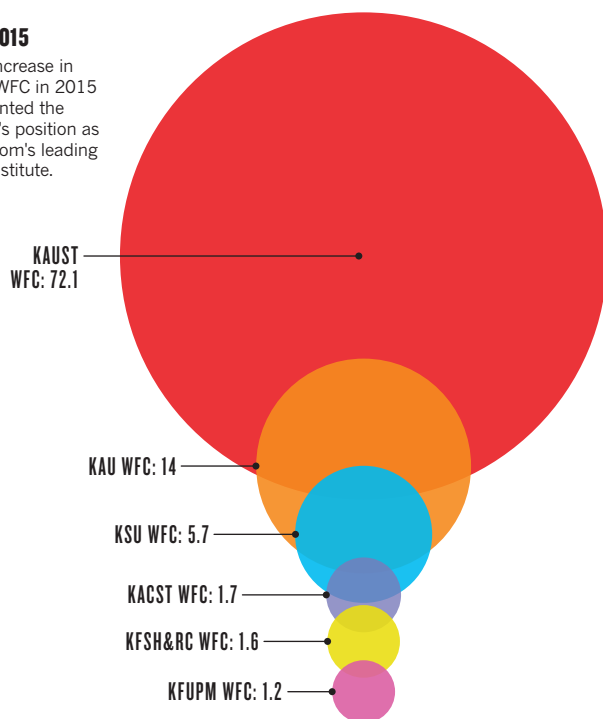
Strong connections with global scientific heavy-hitters and meaningful regional and domestic collaborations have thrust Saudi Arabia into a leading position in the Arab world.

MAPPING GROWTH

Five Saudi Arabian institutes are driving the country's rapid progress in science, with its west-coast institutes leading the way. They have lifted Saudi Arabia eight places higher in the Nature Index from 39 in 2012 to 31 in 2015.

WFC IN 2015

A sharp increase in KAUST's WFC in 2015 has cemented the university's position as the Kingdom's leading science institute.



1. THUAWAL

Located on the Kingdom's west coast, Thuwal is home to KAUST, a graduate-level university with a US\$20 billion endowment, founded in 2009.

2. JEDDAH

A major port on the west coast of Saudi Arabia, Jeddah is home to KAU, one of the fastest rising universities in the Kingdom in the Nature Index.

3. RIYADH

Riyadh is Saudi Arabia's capital and biggest city. It is home to KACST, responsible for putting together the country's science strategy, and KSU, the oldest university in the country.

4. DAMMAM

KFSH&RC, located on the Kingdom's eastern coast in Dammam, has the strongest network of domestic and regional collaborators in the country.

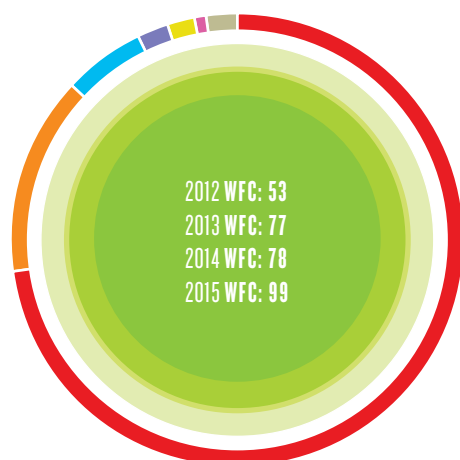
5. DHARRAN

KFUPM in Dhahran has a strong focus on chemistry, and is home to the Dhahran Techno Valley, a business initiative to link research and industry.



SAUDI ARABIA'S RISE

Saudi Arabia's WFC has steadily risen by 85% since 2012, with only a lull in 2014. Nearly 90% of the country's science output in 2015 was driven by KAUST and KAU.

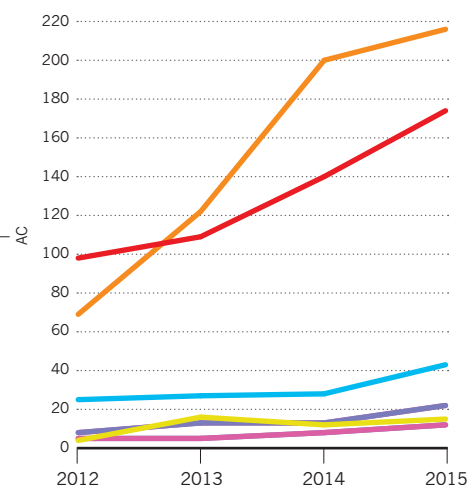


- KAUST: 73%
- KAU: 14%
- KSU: 6%
- KACST: 2%
- KFSH&RC: 2%
- KFUPM: 1%
- Others: 2%

OUTPUT

The AC of Saudi Arabia has risen quickly over the past four years, driven by strong international collaborations.

A country or institution's AC is the number of articles in the index that have at least one author from that country or institution.



LEGEND

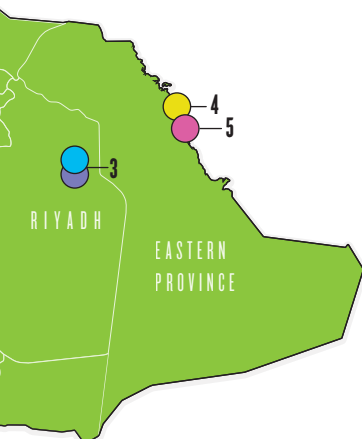
- King Abdullah University of Science & Technology (KAUST)
- King Abdulaziz University (KAU)
- King Faisal Specialist Hospital & Research Centre (KFSH&RC)
- King Saud University (KSU)
- King Abdulaziz City for Science & Technology (KACST)
- King Fahd University of Petroleum & Minerals (KFUPM)

AC: article count
CS: collaboration score
WFC: weighted fractional count

DATA ANALYSIS BY LARISSA KOGLECK

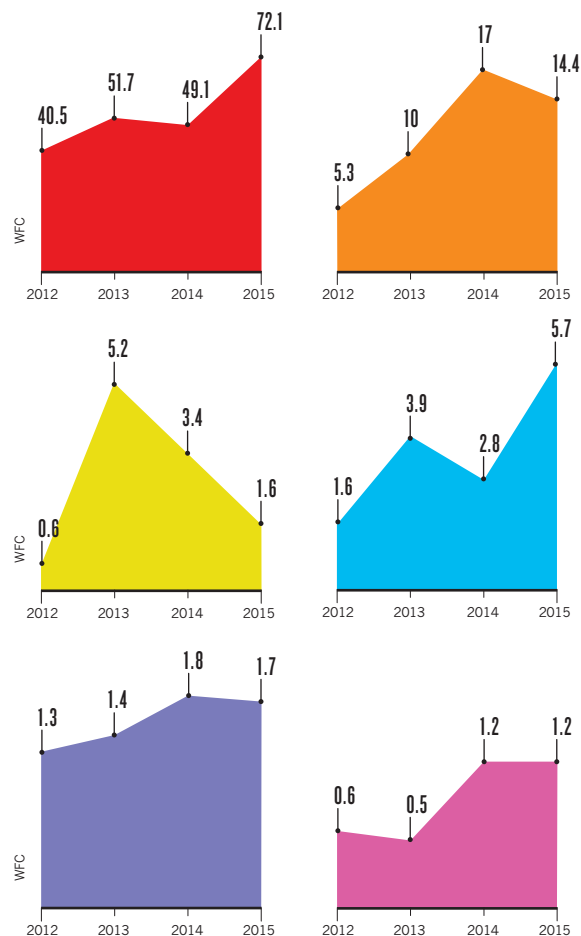
AC

Even though KAUST's WFC is five times higher, KAU leads in the number of articles (AC) published in the Nature Index. Strong international collaborations helped it publish 216 articles in 2015. KSU comes a distant third, with a fifth of KAU's AC.



WFC RISE AND FALL

Most of Saudi Arabia's leading institutes have seen their WFC grow steadily year-on-year since 2012. This has fuelled the country's international standing in the index (see page S13, 'Oiling the wheels on a road to success').

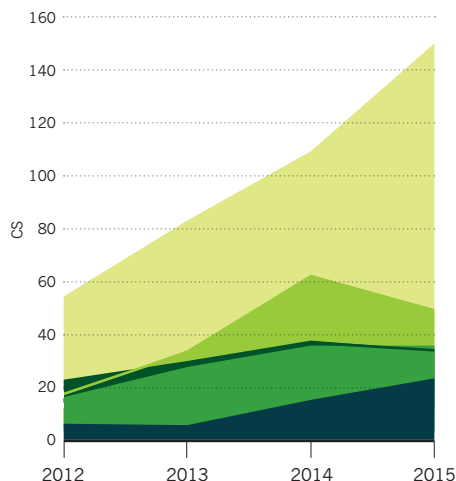


COLLABORATION

Saudi Arabia's top collaborators* have remained mostly unchanged since 2012, with the United States its biggest research partner. Collaborations with China were increasing sharply, but slowed down in 2015.

- United States
- China
- United Kingdom
- Germany
- Canada

*CSs are only for output derived from the bilateral relationship of Saudi Arabia and each partner country.

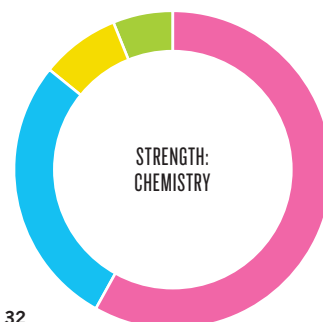


SUBJECT SPLIT*

Chemistry accounts for two-thirds of Saudi Arabia's research in the index.

- Chemistry WFC: 67
- Physical sciences WFC: 32
- Life sciences WFC: 9
- Earth & environmental WFC: 7

*Subjects may overlap. The sum of subject area WFCs may therefore exceed the country's overall WFC.





A view from the Kingdom Tower in Riyadh, the Saudi Arabian capital and a main centre for the country's renewed commitment to science.

A 21ST CENTURY TRANSFORMATION

Saudi Arabia has a bold plan to diversify from its oil industry to create a knowledge economy.

BY SEDEER EL-SHOWK

As the world searches for viable energy sources as alternatives to fossil fuels, Saudi Arabia is striving to diversify in order to secure its future prosperity and reduce its economic reliance on oil. In 2002, the Saudi government established the National Science, Technology and Innovation Policy (NSTIP), an ambitious long-term strategic framework to manage the nation's scientific development and transition to a knowledge-based economy. More than US\$6 billion was allocated to the first phase of the NSTIP, which ran from 2008 to 2014.

The Kingdom's effort to become a knowledge-based economy is spearheaded by its national science agency, the King Abdulaziz City for Science and Technology (KACST), which is responsible for implementing the NSTIP. Ambitious initiatives launched by KACST, such as the Saudi Human Genome Project, have expanded the country's scientific reach.

Saudi Arabia's research landscape has also been transformed by the growth of the King Abdullah University of Science and Technology (KAUST), a graduate-level research university founded on the shores of the Red Sea in 2009, modelled on the structure of Western universities such as Caltech. Mohamed Eddoudi, chair of KAUST's chemical science programme and associate director of its Advanced Membranes and Porous Materials Research Center, believes the university

"What the Kingdom of Saudi Arabia accomplished in the past few years went beyond our expectations. The challenge now is to meet the new goals."

provides an example for other countries in the region, a demonstration that "excellent research can be performed effectively and efficiently anywhere in the world if bright minds are given the world-class tools necessary to be competitive".

The Nature Index

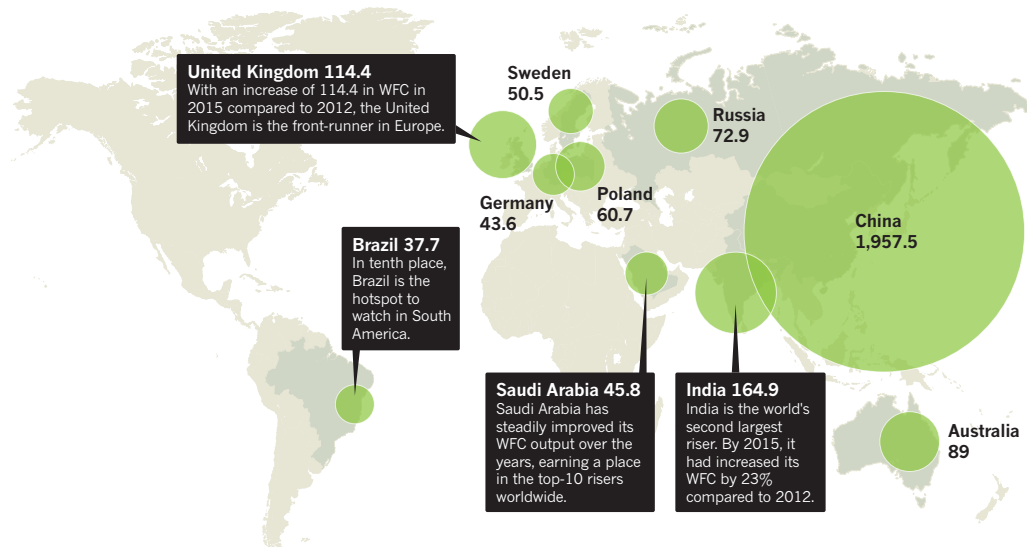
reflects the fruits of these developments. Over the last four years, Saudi Arabia has seen its presence in the index expand rapidly. Its weighted fractional count (WFC), which measures the contribution of authors to scientific papers tracked by the index, has more than doubled, making the Kingdom the eighth largest riser in WFC globally. In 2015, 21 Saudi Arabian institutions were affiliated with authors publishing their research in the Nature Index journals.

NEW FIELDS OF DISCOVERY

Being the world's largest oil exporter, it comes as no surprise that petroleum industries dominate the Saudi Arabian economy, with the oil sector accounting for around half of the country's US\$750 billion GDP and the vast majority of its exports. This focus inevitably instructs the country's research priorities. The bulk of Saudi Arabia's WFC has come from the chemical and physical sciences, which together constitute almost 90% of the country's output in the Nature Index in 2015. In

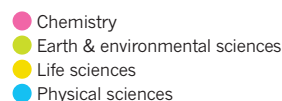
TOP 10 GLOBAL RISERS

This map shows the top 10 countries that have experienced the largest increase of their WFC from 2012 to 2015 worldwide. The circle sizes are relative to this increase. China leads the way having jumped from a WFC of 4,523 in 2012 to a WFC of almost 6,481, an increase of 1,958.



SAUDI ARABIA'S RESEARCH STRENGTHS

Saudi Arabia's overall research output (WFC) in the Nature Index has grown, but different subject areas have experienced varying patterns of growth.



SPOTLIGHT ON LIFE SCIENCES

Often overshadowed by Saudi Arabia's achievements in chemistry and physical sciences, life science research in the Kingdom is beginning to come into its own, with the help of these institutions contributing to its life sciences output in 2015.



particular chemistry has seen a rapid rise over the years, overtaking the physical sciences as the top subject area in 2014 and continuing to expand in 2015. The effort to shift the country towards a knowledge-based economy has capitalized on these strengths, bringing rewards from research projects in advanced material sciences, nanotechnology and photonics.

Despite the country's overt focus on chemistry, it has also been increasing its output in the life sciences as well as in Earth and environmental sciences. Since 2012, contribution to life science papers in the index has doubled, with much of this increase being driven by research at KAUST and the King Faisal Specialist Hospital and Research Centre (KFSH&RC).

KFSH&RC, which focuses almost exclusively on life science research, has been increasing both the number of papers with a KFSH&RC affiliated author as well as the contribution they make. In 2012, its researchers contributed to only four papers that count within Saudi Arabia's article count (AC). This

jumped up to 15 papers in 2015. Its overall contribution, as measured by WFC, has also seen an increase over the years, particularly marked from 2012 to 2013.

"This is due to a strong research platform built of talented researchers, excellent technical capabilities and proper oversight to ensure that our work is done at the highest standards," explains Sultan Al-Sedairy, executive director of KFSH&RC. "The primary propeller was the availability of funding through the NSTIP from KACST." Al-Sedairy also directs the Saudi Human Genome Project, which has driven domestic collaborations and enabled local researchers to publish higher impact papers.

Geneticist Fowzan Alkuraya, a specialist in disease gene discovery, joined KFSH&RC in 2007 after training in the United States. Since then, he has built an extensive network of collaborators throughout the Middle East who connect him with significant patients. "We got access to whole-exome sequencing in 2011, and that's when my lab's gene-discovery pipeline really took off. Instead of discovering

a couple of genes per year, we now discover at least one or two genes per week," he says. Through the Saudi Human Genome Project, his team has free and mostly unlimited access to next-generation sequencing.

KAUST has also seen dramatic growth in its life sciences output, with a more than three-fold increase in its WFC from 2012 to 2015. "It's the nature of the research. Getting results in biology takes several years, and KAUST was only established in 2009, so this is about the earliest you'd expect to start seeing a significant number of publications in high-profile journals," says Pierre Magistretti, dean of KAUST's Division of Biological and Environmental Sciences and Engineering.

Magistretti explains that his division is focusing on a few key areas in order to maximize the impact of its research. While researchers in KAUST's academic divisions are free to pursue basic research, the university also has 11 research centres that focus on applied research in areas of national importance, such as water desalination, desert



Researchers at KAUST'S Advanced Membranes and Porous Materials Research Center, whose work aims for efficient desalination of seawater and wastewater treatment.

agriculture, and solar energy.

In addition to cultivating local talent, KAUST encourages productive collaborations. "It's important that the principal investigators are active and fully based here, but they can enhance their potential with collaborators from outside," says Magistretti. Principal investigators awarded grants through the university's Competitive Research Grants programme can share a percentage of the funds with collaborators if they participate in the grant application, offering an incentive to build international networks.

NURTURING LOCAL TALENT

KAUST's research output in the Kingdom is rivalled only by King Abdulaziz University (KAU), but the two show vastly different patterns of collaboration. Since 2013, KAU-affiliated authors have produced increasingly more articles than KAUST, but its contribution to these papers has remained relatively low by comparison. This may indicate that many of its publications resulted from collaborations in which it played only a small part. KAUST may have contributed to fewer articles, but its overall WFC of 72 in 2015 eclipses KAU's 14. It accounts for 73% of Saudi Arabia's institutional WFC. By pulling their weight in collaborations,

institutes such as KAUST and KFSH&RC not only enhance Saudi Arabia's scientific reputation, they also help build the local talent and capability that is crucial to the Kingdom's goal of a knowledge-based economy.

"It was just a matter of time before we saw local students as lead authors in top-tier journals," says Eddaoudi. "A female Saudi student was the lead author on a high-profile paper my group recently published where 95% of the work was done at KAUST, and the only outside activity was the use of a synchrotron facility in Europe." The paper, published in the *Journal of the*

"More than 60% of the biology students at KAUST are female. And many of them are Saudi."

American Chemical Society in 2015, describes a metal-organic framework which can be used to store methane at room temperature and low pressures, an important step towards the efficient use of the gas as a clean, alternative fuel.

The knowledge and skills that come from scientific training may be particularly beneficial to Saudi Arabian women, who are the subject of heavy cultural restrictions in the Kingdom. "More than 60% of the biology students

at KAUST are female, and many of them are Saudi," says Magistretti. "I think this is something very positive."

The total number of Saudi students studying abroad increased to 200,000 in 2013, says Mansour Alghamdi, director of scientific awareness and publishing at KACST. The Kingdom also benefits from a significant knowledge transfer by students returning home. Alghamdi explains that many return to Saudi Arabia as researchers who continue to publish with their former supervisors, while improving experience levels domestically.

A CHALLENGING TRANSITION

Despite these accomplishments, the transition to a knowledge-based economy has a long way to go, with oil and petroleum-related industries continuing to play the central role in Saudi Arabia. The prioritization of scientific research has yet to be reflected in the country's R&D budget, which was equivalent to only 0.3% of the GDP in 2015 according to a report by Battelle, though the NSTIP calls for an increase to 1.6% of GDP by 2020.

In addition, R&D spending by private firms, though not monitored, is estimated to be very low by international standards, according to Mohammad Khorsheed, the secretary general



ANASTASIA KHRENOVA/KAUST; KACST

KAUST'S Center for Desert Agriculture (left) looks at such possibilities as drought resistant crops and one of the laboratories in KACST's solar village, located 35-km north of Riyadh (right).

of the steering committee for the Saudi Innovation Ecosystem. Khorsheed also highlights social resources as a challenge facing the Kingdom. A knowledge-based economy thrives in a knowledge-based society, yet Khorsheed cites public apathy about science and a lack of interest in education. Only 23 people out of 100,000 work in research and development, and only one in 1,000 people aged 20 to 34 are science and engineering graduates, less than a tenth of the proportion of such graduates in an average population of an EU country. Brain drain, which sees the brightest people of a population leave, also poses a problem, with 25% of science, technology, engineering and mathematics (STEM) graduates emigrating annually.

But, officials remain optimistic. Alghamdi does not think a lack of human resources will be a significant issue due to Saudi Arabia's "unprecedented expansion in higher education in the last few years," going on to cite OECD predictions that the Kingdom will see a six-fold increase in tertiary degrees by 2030.

KAUST's gifted student programme identifies promising students in the penultimate year of high school to spend a summer at KAUST and then travel to the United States on a KAUST fellowship after they graduate. The students spend a foundation year learning

about the American system before applying to top universities, and after completing their undergraduate studies return to KAUST for graduate studies. "It's a valuable programme and could produce remarkably qualified students," says Magistretti.

TOWARDS A KNOWLEDGE-BASED ECONOMY

With strong financial support and young scientists like Alkuraya returning from abroad, there's little doubt that Saudi universities will continue to improve their research output as labs mature.

The first phase of the NSTIP saw Saudi researchers establish local infrastructure while engaging in collaborations worldwide. "Some of the collaborations were not very successful, but there are lessons to be learned from them. Besides transferring technology to the

"Some of the collaborations were not very successful, but there are lessons to be learned from them."

Kingdom, there are also serious efforts to transfer technologies from academia to the industrial sector, though it's a big challenge," says Khorsheed.

Khorsheed is helping address that

challenge by building a framework to support innovation in Saudi Arabia. He notes that despite their abundant liquidity, Saudi investors tend to be risk-averse and reluctant to put funds into young, technology-driven companies.

In an effort to overcome this, KACST has established the BADIR Program for Technology Incubators, which aims to encourage innovative startups. Several companies have already graduated from the incubator programme, including a company which developed a surgical dressing for diabetic foot ulcers out of an industrial waste product from prawn shells.

The second phase of the NSTIP, which began in 2015 and runs until 2019, calls for Saudi Arabia to build on the developments in its national infrastructure and capabilities to make the country a regional leader in science, technology and innovation. This will require a continued focus on research and patent production, coupled with a strong programme to transfer technology to the private sector and an intensive effort to develop the Kingdom's human resources.

"What the Kingdom of Saudi Arabia accomplished in the past few years went beyond our expectations," says Alghamdi. "The challenge now is to meet the new goals of the coming stages."

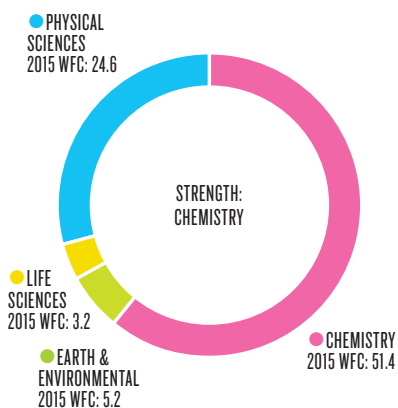
KAUST

2015 WFC: **72.06** 2015 AC: **174**2015 NATURE INDEX GLOBAL RANKING: **174**

COMING OF AGE

The King Abdullah University of Science and Technology (KAUST) was established by the late King Abdullah in 2009 with an endowment of US\$20 billion to provide a state-of-the-art institution for skilled, passionate scientists to carry out cutting-edge research. The university was envisioned as a modern-day House of Wisdom, the influential intellectual centre of the Islamic Golden Age (which ran from the ninth until the thirteenth century).

Located on the Red Sea coast, the campus includes extraordinary core facilities which provide the infrastructure necessary for world-class research. "Scientists here at KAUST are expected to do something that makes a difference," explains Pierre Magistretti. Principal investigators are encouraged to secure grants, but are also offered generous baseline funding to give them the freedom to pursue original research that could result in breakthrough discoveries.



*Subjects may overlap. The sum of subject area WFCs may therefore exceed the institution's overall WFC.

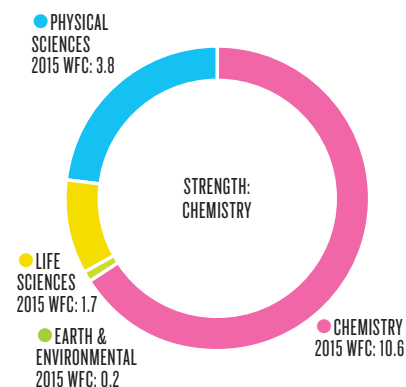
KAU

2015 WFC: **14.43** 2015 AC: **216**2015 NATURE INDEX GLOBAL RANKING: **601**

KAU campuses are single-gender.

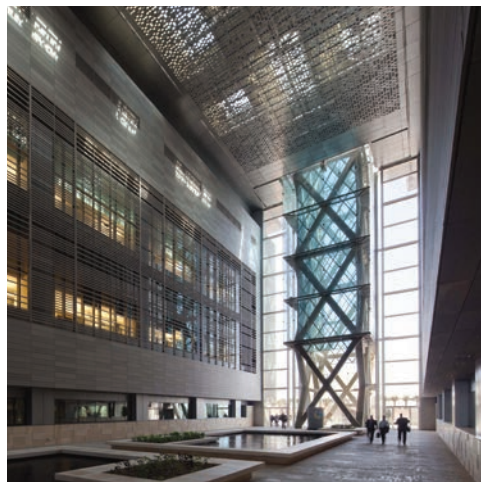
BUILDING BRIDGES

King Abdulaziz University is one of the leading institutes in the Kingdom, with more than 80,000 students at its single-gender campuses. KAU nearly tripled its total output, as measured by WFC, in journals in the index between 2012 and 2015, however, it still heavily depends on external collaborations which have continued strongly over the past few years.



"Our Distinguished Scientists programme has enabled hundreds of international collaborators to visit, research and lecture, increasing collaboration with universities worldwide," explains Adel Alahmadi, KAU's general supervisor for scientific affairs. In future the university wants to increase funding directed at local cooperation, while remaining focused on high-impact research. "Solving local problems will be given a higher priority, with research groups led by prominent local researchers getting stronger support."

KAUST



KAUST



KAUST's core labs on its modern campus on the Red Sea, which acts as a natural lab for marine studies.

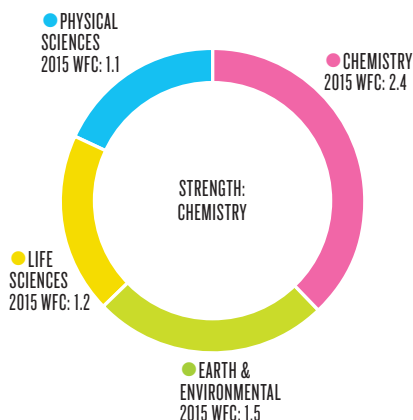
KSU2015 WFC: **5.7** 2015 AC: **43**2015 NATURE INDEX GLOBAL RANKING: **1,025**

KARL-JOSEF HILDENBRAND/DPA/PA IMAGES

King Saud University in Riyadh.

A LOCAL LEADER

Established in 1957, King Saud University was Saudi Arabia's first university. It accounts for about a quarter of the country's scientific output and, until 2014, produced more papers annually than any other institute in the Kingdom, when it was overtaken by King Abdulaziz University. Despite its leading role in the country, KSU ranks behind



KAUST and KAU in the Nature Index because of the focus of its research strategy.

"In support of the NSTIP, KSU targets key research areas which involve use-inspired research, and these areas generally do not conform to the list of journals tracked by the Nature Index," explains Rshood Khraif, KSU's dean of scientific research. "In addition, much of our research output is produced as articles or books published in Arabic, which don't get indexed by the Web of Science."

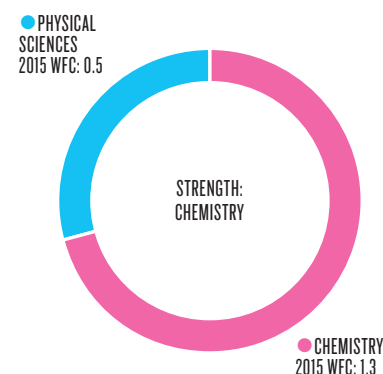
KACST2015 WFC: **1.71** 2015 AC: **22**2015 NATURE INDEX GLOBAL RANKING: **1,732**

KACST

An impression of KACST's Maarifah centre.

A GUIDING FORCE

The King Abdulaziz City for Science and Technology (KACST) serves as the Kingdom's national laboratories and its science agency. While its direct output accounts for only a small fraction of Saudi Arabia's AC, it plays a central role in coordinating and facilitating research across the nation. KACST is responsible for managing the country's science policy, funding research, and building and



maintaining the infrastructure to support scientific research.

"We identified what the most important areas of research for Saudi Arabia are, and plan how we need to address each of these to reach a knowledge economy," says Abdulaziz Al-Swailem, vice-president for scientific research support at KACST.

It has established research centres focused on a wide range of topics, as well as major initiatives such as the Saudi Human Genome Project. It also promotes commercialization of research via projects such as the BADIR Program for Technology Incubators through its Technology Development Center.

KFSH&RC2015 WFC: **1.64** 2015 AC: **15**2015 NATURE INDEX GLOBAL RANKING: **1,762**

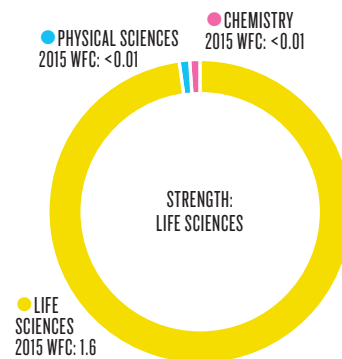
EPA/ALAMY

King Faisal Specialist Hospital in Jeddah.

GENES AND GENOMES

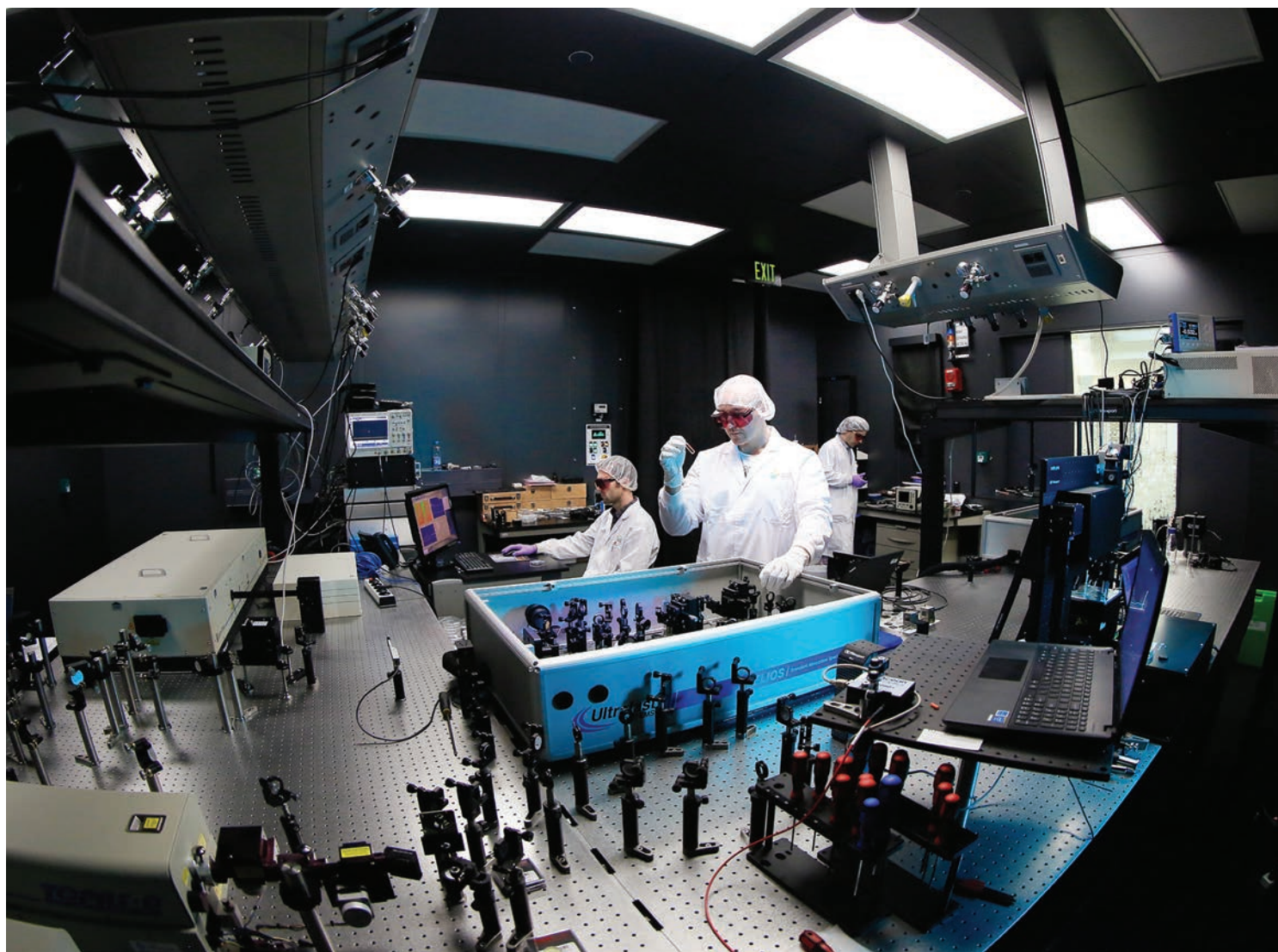
Established in 1975, the King Faisal Specialist Hospital and Research Centre is one of Saudi Arabia's leading medical and academic institutes. It serves as the national referral and research centre for oncology, organ transplants, cardiovascular conditions and genetic diseases.

KFSH&RC is also home to one of the Saudi Human Genome Project's high-throughput sequencing labs. More than



half of marriages in Saudi Arabia are between relatives, which results in high rates of genetic disease in the Kingdom. This prevalence provides a backdrop of rich research opportunity to discover the genes behind these diseases. Researchers at KFSH&RC have created a 'gene panel' which tests 3,000 genes to identify genetic disorders in patients.

"There's a very strong clinical and research interaction," says Sultan Al-Sedairy. "As a result, we have an elite institution researching unique pathologies with the capability to provide samples that are clinically well-defined and then dissect them genotypically." ■



The Solar & Photovoltaics Engineering Research Center at KAUST is among the university's many endeavours aimed at finding alternative energy sources.

MAKING THE MOST OF FINANCIAL MIGHT

In a troubled region, Saudi Arabia is capitalizing on its relative stability and resource wealth.

BY PAKINAM AMER

In a region marred by ongoing conflict, few nations in the Middle East have the financial and political capacity to prioritize expanding the frontiers of science, however, Saudi Arabia is tapping into its large stream of oil revenue to fund a research revival with the aim of becoming a regional science leader.

Over the past decade, the Saudi Arabian research environment has changed dramatically, not just in terms of greater funding, but in the way research is conducted. The Kingdom has built state-of-the-art research institutes, forged relationships with renowned foreign institutions and developed a visionary science strategy that extends into 2030.

This vision is now starting to bear fruit.

Saudi Arabia had the fastest growth in weighted fractional count (WFC) output in the Middle East, surpassing all the region's other countries. Since 2012, its WFC has more than doubled, from 46 to 99 by 2015.

AN ARAB LEADER

The oil-rich Kingdom is not facing very robust competition from its Arab states neighbours. In the United Arab Emirates (UAE) and Qatar, research is indeed growing and branches of well-established Western universities are setting up, but the UAE's position is a very distant second with a WFC of 12.

Though Egypt has been the regional leader in scientific research, its recent political turmoil has hampered growth in science. The region's former stronghold has experienced a

decline in WFC and slipped behind the UAE to third in the Arab world with a WFC of 9 in 2015.

Even new high-calibre institutions like Zewail City of Science and Technology, and established powerhouses like Cairo University have not been able to keep up with the plethora of research opportunities, grants and partnerships in Saudi Arabia.

"We've set 25-year goals to create a regional and international impact in research."

Elsewhere in the Arab world, conflict thwarted opportunities for countries to sponsor scientific endeavours, with many losing the infrastructure, funding and expertise required to build research

enterprises or provide an adequate level of basic education.

In Saudi Arabia, by contrast, according to the British Foreign and Commonwealth Office's research and analysis department, education was allocated 25% of total budget expenditure in 2015, the equivalent of more than US\$54 billion, which included the construction of three new universities in Jeddah, Bisha, and Hafr Al-Baten, major refurbishment of existing universities and upgrading of hundreds of new schools.

REGIONAL CHALLENGERS

Beyond its immediate neighbours, Saudi Arabia has more serious regional competition — mainly in Turkey and Iran — but the country is forging ahead in several fields.

In 2015, Saudi Arabia has again overtaken the two countries, making it the leader of the three in terms of WFC output in the index.

“Beyond its immediate neighbours, Saudi Arabia has more serious regional competition.”

With a WFC around 45% bigger than either Iran's or Turkey's, Saudi Arabia is also the most prolific among the three in the total number of papers published in journals in the index, having contributed to 479 articles in 2015.

Saudi Arabia is comfortably ahead in terms of growth, with an 87% increase in WFC from 2012 to 2015, thereby steadily rising faster than Iran, which has seen a 4% decline, and Turkey, which grew by 8%, over the same period. This rise may be an indication of the country already beginning to reap the reward of its investment.

It therefore comes as no surprise that Saudi Arabia's institutions are the front-runners amongst regional competitors. King Abdullah University of Science and Technology (KAUST) and King Abdulaziz University (KAU) are leading the top-10 institutions in the three countries, the rest of which are made up of three universities from Turkey and five institutions from Iran. No other institutes in any Arab state make the list.

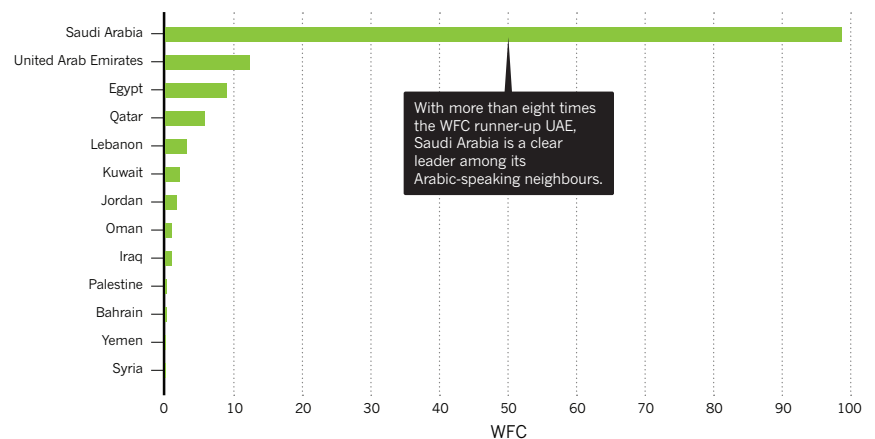
RESEARCH PRIORITIES

The specific areas of research pursued by the Kingdom are part of the country's National Science, Technology and Innovation Plan (NSTIP), formulated in 2008, and set to instruct the country's research until 2030.

Since then, R&D in Saudi Arabia has grown sharply, according to Mansour Al-Ghamdi, director of the general directorate of scientific awareness and publishing at King Abdulaziz City of Science and Technology (KACST). He expects this trend to continue steadily over coming years. The first stage of the plan aims to position Saudi Arabia as a regional leader, and the Kingdom has taken strides over the past four years towards that ambition, according to

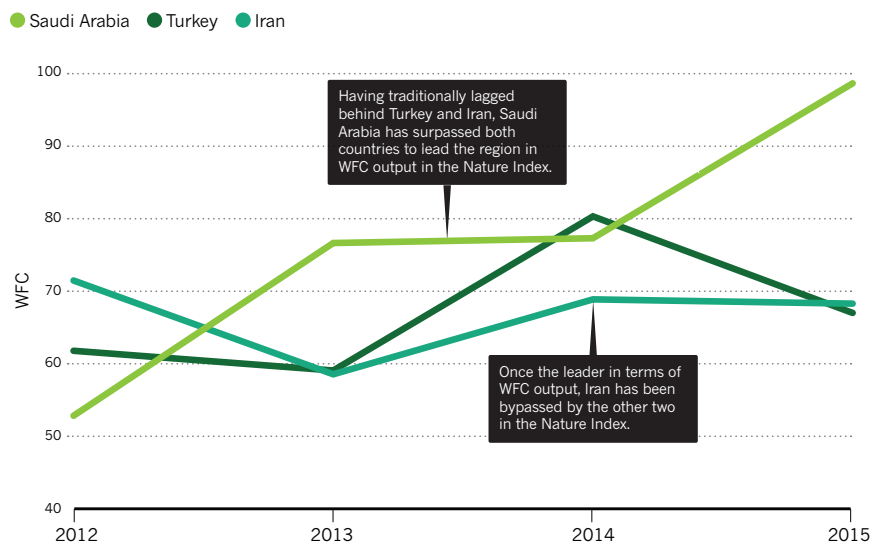
ARAB WORLD IN THE NATURE INDEX

This bar chart shows the 2015 overall output (WFC) of Saudi Arabia and its Arabic-speaking neighbours on and around the Arabian peninsula.



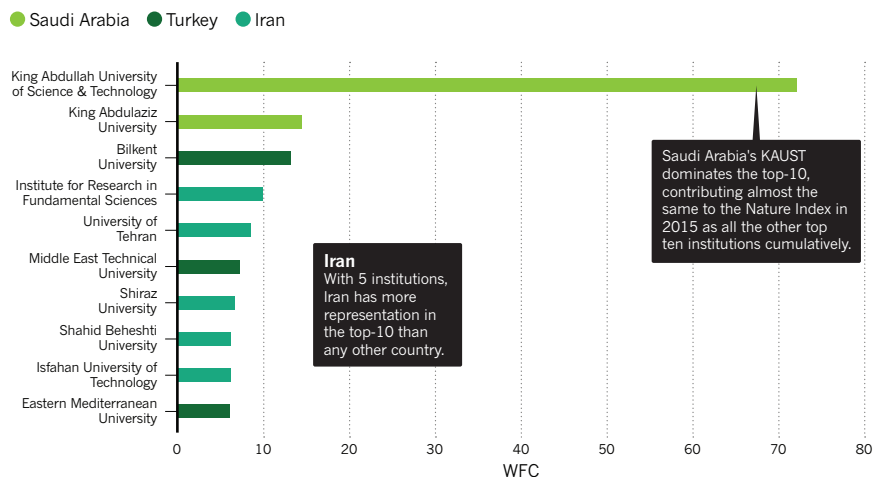
RISE AND FALL OF REGIONAL POWERHOUSES

The traditional regional stalwarts of science, Iran, Turkey and Saudi Arabia have seen a shift in their relative positions in the Nature Index over the years. Below is a timeline of their WFC output since 2012.



FLYING HIGH

Iran, Turkey and Saudi Arabia have been viewed as scientific leaders in the region. Below are the three countries' top-10 institutions by WFC in the Nature Index for 2015.





The Upstream Petroleum Engineering Research Center at KAUST (left) and mechanical engineers and plant scientists at work at KAU (right).

the index rankings.

“We’ve set goals over 25 years to make a regional and international impact in research, but instead of funding everything, we’ve specified 13 priority areas,” says Nasser Al-Aqeeli, dean of research at King Fahd University of Petroleum & Minerals (KFUPM).

These priority areas span several technological fields, explains Al-Aqeeli, and the plan is to pour the bulk of the huge funding reserves Saudi Arabia is allocating to research into these priority disciplines to help make a difference. Saudi Arabia’s enormous oil industry clearly drives the country’s science focus. The country’s science output in chemistry is highest and, between 2012 and 2015, Saudi Arabia’s chemistry WFC experienced a 190% growth, the largest increase for any country in the Middle East in any subject area tracked by the Nature Index. Its 2015 WFC of 67 in chemistry places it well ahead of Turkey, Iran, Egypt and other Arab neighbours.

“Publishing in chemistry is easier than in other fields,” says Al-Aqeeli. “It’s grabbing the attention of the industry here, so by default funding it is easier.” Materials science is also receiving increased funding, he adds.

Research in physical sciences in Saudi Arabia is second only to chemistry. Its output has

been fluctuating, however, with its WFC hovering around 30. In 2015, Saudi Arabia’s WFC of 32 has been outranked by Turkey, with a WFC of 34, and Iran, with a WFC of 53.

Research output in the life sciences also grew rapidly over the past four years, doubling from 2012 to a WFC of 8.5 in 2015. However, compared to its 102 papers published in this discipline, this is the smallest relative author contribution of all the four subject areas, with only 8% per paper — a trend seen in other countries in the region.

Earth and environmental sciences, which has a low uptake by institutions in the region, has also seen a recent growth spurt. After three years of fluctuation, Saudi researchers have increased the country’s output in the field by 153% in 2015 compared to 2012, giving Saudi Arabia a slight edge over Turkey.

At KAUST Earth science is still a nascent field, according to J. Carlos Santamarina, professor of Earth science and engineering at the university, but it is growing.

“Chemistry grabs the attention of industry here, so by default, funding it is easier.”

Researchers at KAUST are exploring a variety of topics including deep Red Sea currents and rifts, volcanology and seismic activity, and dust and atmospheric science.

“Clearly, these are all critical areas in Saudi Arabia,” he says. “And, knowledge generated in these areas is relevant worldwide.”

The research priorities of the Kingdom are shaped by its needs and its attempts to best exploit its natural resources, while keeping an eye on emerging alternative energy sources. The next phase of the NSTIP is also focused on linking research to industry.

“We’ve been doing good work in terms of executing research but it was very difficult to translate patents and discoveries to economical values. Many patents failed to materialize,” says Al-Aqeeli. This difficulty is why the Kingdom continues to focus on specific research areas seen as being the most likely to be translated into ventures. It explains the focus on chemistry and applied physical sciences, and a lack of enthusiasm for other areas such as life sciences.

Instead of throwing money at all types of research and seeing what sticks, Saudi Arabia is opting for a process of thoughtful selection. “We’re trying to redefine excellence when it comes to research,” says Al-Aqeeli. ■



KAUST students embark on a new school year with a commencement ceremony. The relatively new university has quickly made an impact on the Nature Index.

OILING THE WHEELS ON A ROAD TO SUCCESS

With the benefit of a sustainable plan and the funds to back it, Saudi Arabia is aiming high.

BY PAKINAM AMER

Saudi Arabia's scientific development may be in its infancy, but the oil-rich Kingdom is making strides in terms of research investment and publication — with a clear ambition to one day join those in the highest echelons.

In 2012, Saudi Arabia had a weighted fractional count (WFC) of 52.84 in the index, sitting behind Turkey, Iran, Mexico, Chile and South Africa. In four years it rose 86.8% to reach a WFC of 98.67, leapfrogging all these countries to compete with Chile and Argentina globally. Saudi Arabia ranks at number 31 in the world in terms of WFC — up from 39 in 2012.

The country has risen even higher in specific subject areas. In chemistry, for example, it has surpassed countries with a strong scientific impact like Finland and Ireland,

with its WFC rising to 66.54, achieving almost a three-fold increase from its position in 2012.

Institutionally, the country's leading science hub King Abdullah University of Science and Technology (KAUST) made an impressive leap in its WFC between 2012 and 2015, carving a place for itself to compete with American and European research powerhouses.

In just four years, its WFC has risen to become higher than those of prestigious institutions including the European Organization for Nuclear

"Its rise up the ranks depends on a 'self-correcting mechanism' of a slow start to sustainable growth."

Research (CERN), Brookhaven National Laboratory (BNL), the University of Georgia, United States, and Dresden University of Technology, Germany, to name a few. The

output of all of these institutions dwarfed KAUST's in 2012, but KAUST's impressive trajectory since then has seen its WFC shoot to 72 in 2015, overtaking these heavy-hitters.

The country's science development ambitions have been backed by action. Since 2008, the country has embarked on a multi-tiered strategy that will see the Kingdom overhaul its science infrastructure, build high-spec labs, secure grants for research in priority areas in applied science, and link science to industries that drive the economy.

The strategy, broken into four stages to be implemented by 2030, aims to eventually "see Saudi Arabia become a leader in Asia and give it an economic power based on science," says Abdulaziz Al-Swailem, vice president of scientific research support at King Abdulaziz City for Science and Technology (KACST).

The Kingdom's science investments focus on applied research that feeds directly into the



The Saudi Human Genome Project will sequence 100,000 human genomes to conduct biomedical research in the Saudi population.

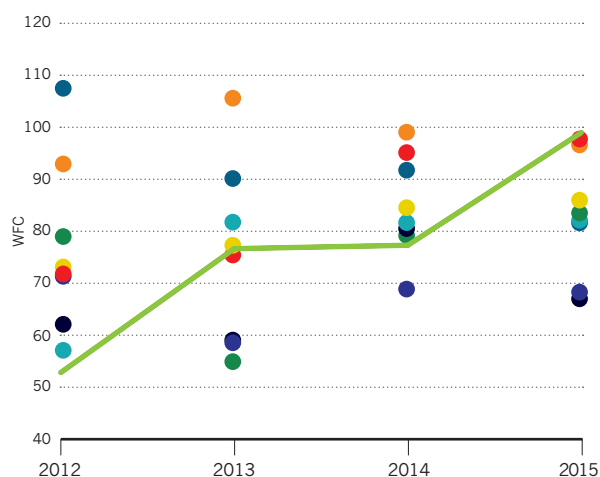
SAUDI ARABIA'S MARCH TO THE TOP

Saudi Arabia's efforts to boost its scientific research have been paying off, with its output in the Nature Index (WFC) rising steadily over the years. The two graphs below highlight Saudi Arabia's rise compared to other nations, both overall and for chemistry.

OVERALL OUTPUT

In 2012 Saudi Arabia's overall output in the index was below all the countries shown, but continuous efforts have seen the Kingdom's WFC rise to overtake them all in 2015.

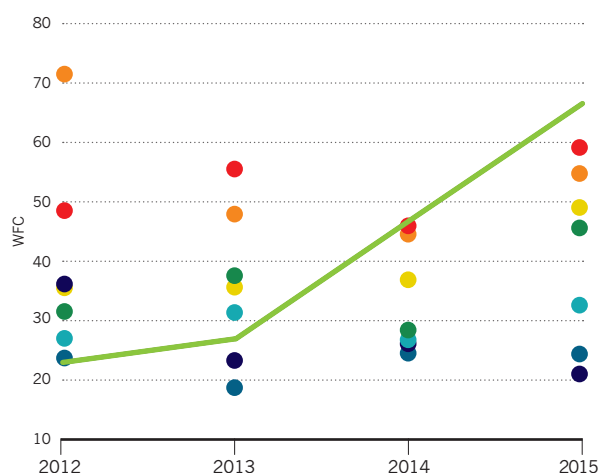
- Saudi Arabia
- Chile
- Argentina
- Mexico
- Hungary
- South Africa
- Greece
- Iran
- Turkey



CHEMISTRY

More marked than its overall rise, Saudi Arabia has made great strides in chemistry. After accelerated growth, which saw the Kingdom's chemistry WFC triple since 2012, it has outshone many larger players in the field in 2015.

- Saudi Arabia
- Finland
- Ireland
- Brazil
- Portugal
- New Zealand
- Turkey
- Greece



country's industrial interests, particularly the oil and energy sector. But even in its strong subjects, chemistry and the physical sciences, Saudi Arabia's WFC remains modest compared to big players in Asia like China, Japan and South Korea.

To truly swim comfortably with these bigger fish, Saudi Arabia may benefit from looking at successful emerging economies in Asia.

One inspiration could be India. In addition to multi-disciplinary scientific and technical advancements that have improved its output in the index from 736.5 to 901.4 in the past four years, the subcontinental giant has joined the exclusive club of countries that have launched successful space missions.

Like Saudi Arabia, India's leading research institutes focus on chemistry, and their total output currently outstrips their Saudi Arabian counterparts by almost a factor of seven (the latter surpassing 472 in 2015, while the former is 66.5).

India's prowess in chemistry is something that Saudi Arabia can aspire to, considering that working conditions for researchers in the Kingdom are more conducive.

India's science ecosystem is far from perfect. Research funding cannot keep up with inflation and a general slowdown in the country's economy. In addition, commentators from the research community say the funding processes are lengthy, bureaucratic, and provide little feedback when applications for grants are turned down. Meanwhile, Saudi Arabia's healthy stream of oil revenue provides assured funding for the country's state-of-the-art research facilities.

While India has slightly increased spending and dedicated US\$1.19 billion for the next fiscal year (2016–2017) for science, it has around 700 universities and 200,000 full-time researchers drawing on the same funding pot. By contrast, Saudi Arabia has pledged an education and training budget of US\$50.9 billion for next year, which includes higher education and scientific research. With a total population of just 30 million, it has a much lower number of full-time researchers competing for the available resources.

Another impressive trajectory that Saudi Arabia might look to emulate is that of Singapore, which has a smaller population as well and has managed to climb high in the index. Like the Kingdom, Singapore also has a focus on chemistry research, and it has put together a similar top-down national science strategy for research institutes across the country. Both countries have strong collaborations with top universities around the world and are welcoming of foreign researchers in their efforts to drive innovation.

Mansour Alghamdi, director of the general directorate of scientific awareness and publishing at KACST, is optimistic that Saudi Arabia can bridge the large gap that currently exists in the volume of scientific

output between it and such countries as India and Singapore.

“The Kingdom of Saudi Arabia has a clear plan to do so and it has the resources,” he says.

FUTURE GROWTH

In 2012, Saudi's ranking in research output, with a WFC of 52.8, meant it was comparable with countries like South Africa, Turkey and Iran, all hovering around the 60–70 mark. Its WFC stood way below countries like Mexico, Hungary, Chile, Greece and Argentina.

Four years later, the country's research outlook is very different and it is surpassing countries like Argentina, Mexico and Hungary in the index, and levelling the playing field with Chile. Chemistry research led the country's rapid rise to surpass these countries, but its life sciences and physical sciences WFCs of 8.5 and 31.5 still lag behind.

However, the Kingdom's AC has been steadily growing in these two fields over the past four years, hinting at the ever-increasing significance of international collaborations. It seems that Saudi Arabian researchers are casting their nets ever wider and are participating in publishing more articles, to the detriment of the WFC accredited for these articles.

Though international collaboration has proved fruitful, Saudi Arabia must keep a focus on nurturing home-grown talent, says Nasser Al-Aqeeli, dean of research at King Fahd University of Petroleum & Minerals (KFUPM), based in Dhahran's ‘techno valley’ in the eastern region of the Kingdom. In the next five years, he says, the country will focus on a programme for national capacity building.

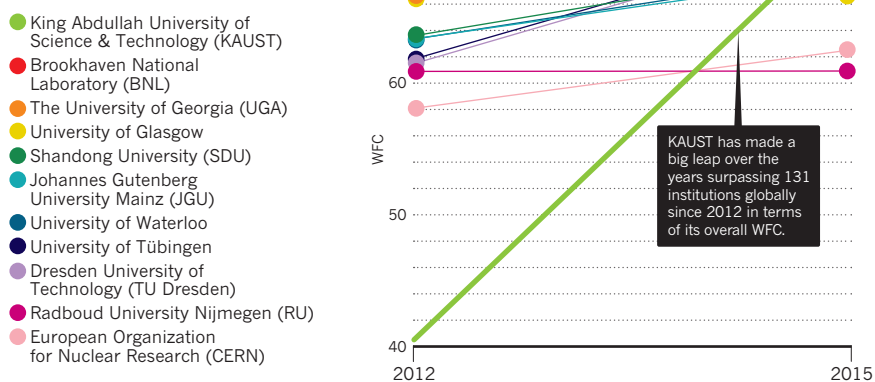
“Saudi Arabia could look to some successful emerging economies for inspiration.”

A good first step was the Saudi government's decision to create a large scholarship programme in 2005, arguably the largest in the world, which has seen more than 200,000 young Saudi Arabians studying abroad. This makes Saudi Arabian students in the United States the fourth largest bloc of expatriate students, following those of China, India and South Korea. The government hopes these students will come back and drive a scientific culture in the country.

Saudi Arabia is also looking to increase its applied research focus, which is an integral part of the current phase of its national science strategy, while securing good funding for basic research as well. Al-Aqeeli says that Saudi's journey involves what he termed a “self-correcting mechanism” where the country is having a slow start in high-impact research, but a more sustainable one. An eventual future move towards basic research might help Saudi Arabia's research capacity to mature. ■

AN INTERNATIONALLY RISING STAR

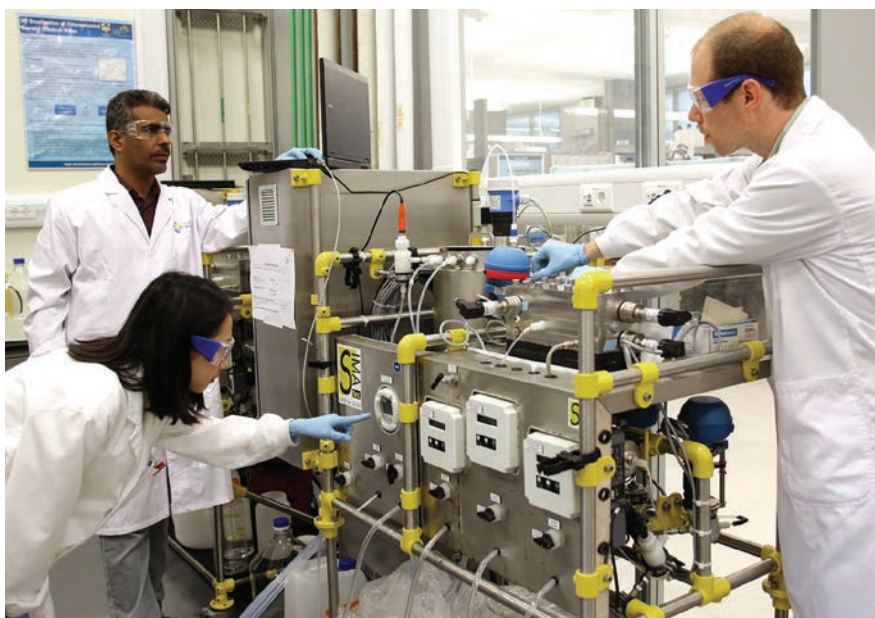
This graph shows KAUST's rise compared to a selection of other institutions*.



*Institutions shown are those that were furthest above KAUST in 2012, have experienced overall growth in WFC by 2015 and have been overtaken by KAUST in 2015. For clarity, only 2012 and 2015 data points are shown.



KAUST



KAUST

Saudi Arabian researchers benefit from cutting-edge labs and generous funding that has boosted the country's R&D.



Welcoming international researchers, Saudi Arabia has forged worldwide collaborations that helped rapidly boost the country's science output.

SHARED KNOWLEDGE IS KEY TO A KINGDOM

International collaboration is yielding major breakthroughs and an increase in quality output.

BY NADIA EL-AWADY

Institutions in Saudi Arabia are casting their nets far in search of collaborative research partners. In 2015 scientists affiliated with Saudi institutions co-authored papers with counterparts from 89 countries in journals included in the Nature Index.

The bulk of these collaborations are with global research powerhouses, rather than with close regional neighbours. The three countries with which Saudi Arabia collaborated most between 2012 and 2015 are the United States, China and — after overtaking Germany in 2015 — the United Kingdom. In the index, Saudi collaborations with all of its top 10 international partners have increased in recent years. Joint research with the US grew

the most, but Saudi research outputs with China have also nearly tripled during that period, as measured by collaboration score, which tallies the sum of all of the Kingdom's bilateral research collaborations.

Saudi Arabia's growing involvement in international collaboration follows its growth in overall output in the Nature Index, in particular in chemistry and the physical sciences. The country's favoured collaborators don't always follow the broader pattern when subject areas are considered in detail. In chemistry, for example, Germany is still its second-largest collaborator after the United States, ahead of China and Canada.

LEADING INSTITUTIONS, DIFFERENT PATTERNS

The two key players in Saudi Arabia's rising international collaborations are King

Abdulaziz University (KAU) on Saudi Arabia's west coast in Jeddah and its closest competitor, King Abdullah University of Science and Technology (KAUST), located about 135km north of Jeddah in Thuwal.

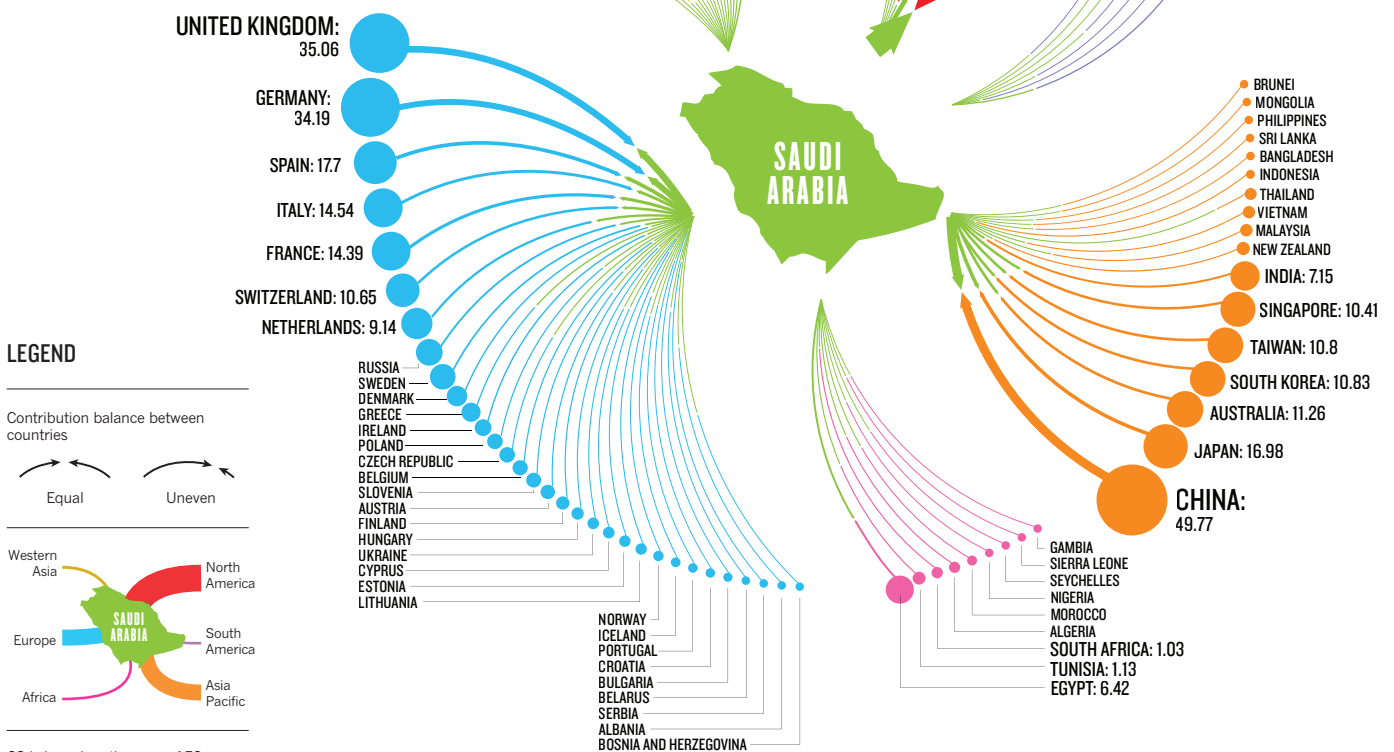
KAU collaborations with US institutions made up 49% of all collaboration scores between Saudi Arabia and US institutions in 2015, while its collaborations with Chinese and UK institutions represented 54% and 21% of collaboration scores between Saudi

institutions and their counterparts in these countries. Collaborations involving KAUST, meanwhile, represent 27%, 23% and 44% of collaboration

"It has to be complementary. You don't want someone riding on the back of somebody else."

INTERNATIONAL COLLABORATIONS

In 2015 Saudi researchers worked with counterparts in 89 countries to produce research counted in the Nature Index. This network diagram shows Saudi Arabia's collaborations with other countries in the index in 2015. The circle sizes relate to collaboration score, which is shown for Saudi Arabia's top collaborators in each region.



CS is based on the sum of FC resulting from collaborative papers between Saudi Arabia and its partner country.

scores with US, Chinese and UK institutions for that same year. Analysing contributions of these two universities to their collaborations reveals distinct patterns, however. Between 2012 and 2015, KAU had a larger collaboration score than KAUST, but its own contribution to the collaborative efforts remains lower.

A factor that could contribute to this may be multiple affiliations on papers. When authors cite multiple affiliations on papers, the Nature Index divides credit among the affiliated institutions through the fractional count (FC) measure. The more affiliations an author has, the smaller the FC attributed to each institution. Collaboration scores for an individual institution are therefore reduced.

KAUST has specifically aimed at attracting international faculty to its campus. "When we hire people, we really look for commitment,"

says Jean Fréchet, KAUST's vice president for research. "We want them to understand that we are looking for goals of excellence. They have a great research environment here, so generally when we hire people we hire them to come full-time."

WORKING WITH THE BEST

When KAUST was founded in 2009, it set up a global collaborations programme with several international universities to help it get established, hire researchers and build laboratories. This programme, which ended in 2015, is the reason KAUST shows major collaborations in the index with France, Singapore, the US and the UK, explains Fréchet. "But it's not a sustainable model in the long term," he says. "This was entirely driven from the top. It had advice from academics, but we had nobody

to drive the programme. Now we are an ongoing institution and we are trying to make sure that our researchers can choose who they work with," he says.

Fréchet says KAUST researchers are encouraged to work with the best experts in their fields. "It has to be complementary. You don't want to have somebody riding on the back of somebody else," he says. "In a collaboration, both parties have to provide something and you really want complementary expertise."

This strategy seems to have paid off. When it comes to co-authoring papers in the Nature Index, KAUST has consistently been contributing roughly the same collaboration scores as its international partners.

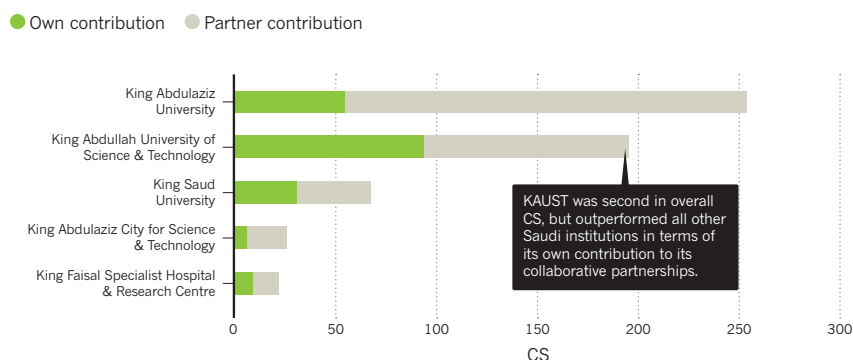
KAUST researchers are encouraged to spend up to 40% of their baseline research funding on external collaborations, says Fréchet. The



Many of KAUST's international collaborations address regional challenges, such as synthetic membranes research for water purification at the Advanced Membranes and Porous Materials Center.

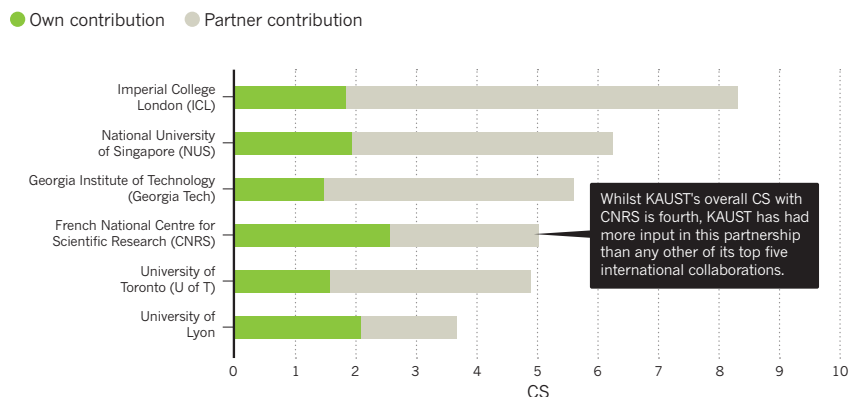
SAUDI ARABIA'S TOP COLLABORATING INSTITUTIONS

KAU led in terms of overall CS in 2015. The bar graph shows Saudi Arabia's top five institutions in 2015 by CS.



KAUST'S TOP FIVE INTERNATIONAL PARTNERSHIPS

KAUST has formed bilateral collaborations with 414 international partners in 2015.



university's research centres are also encouraged to spend 20% of their budgets externally to bring expertise that may be lacking at the young university.

Under its new collaboration model, KAUST is now funding six highly multidisciplinary programmes on sensors research. "We are not only setting up a worthwhile scientific programme," says Fréchet, "but we are helping to broaden our own people. We make them think out of the box. We make them think about something they have never thought about before, because it's not in their field."

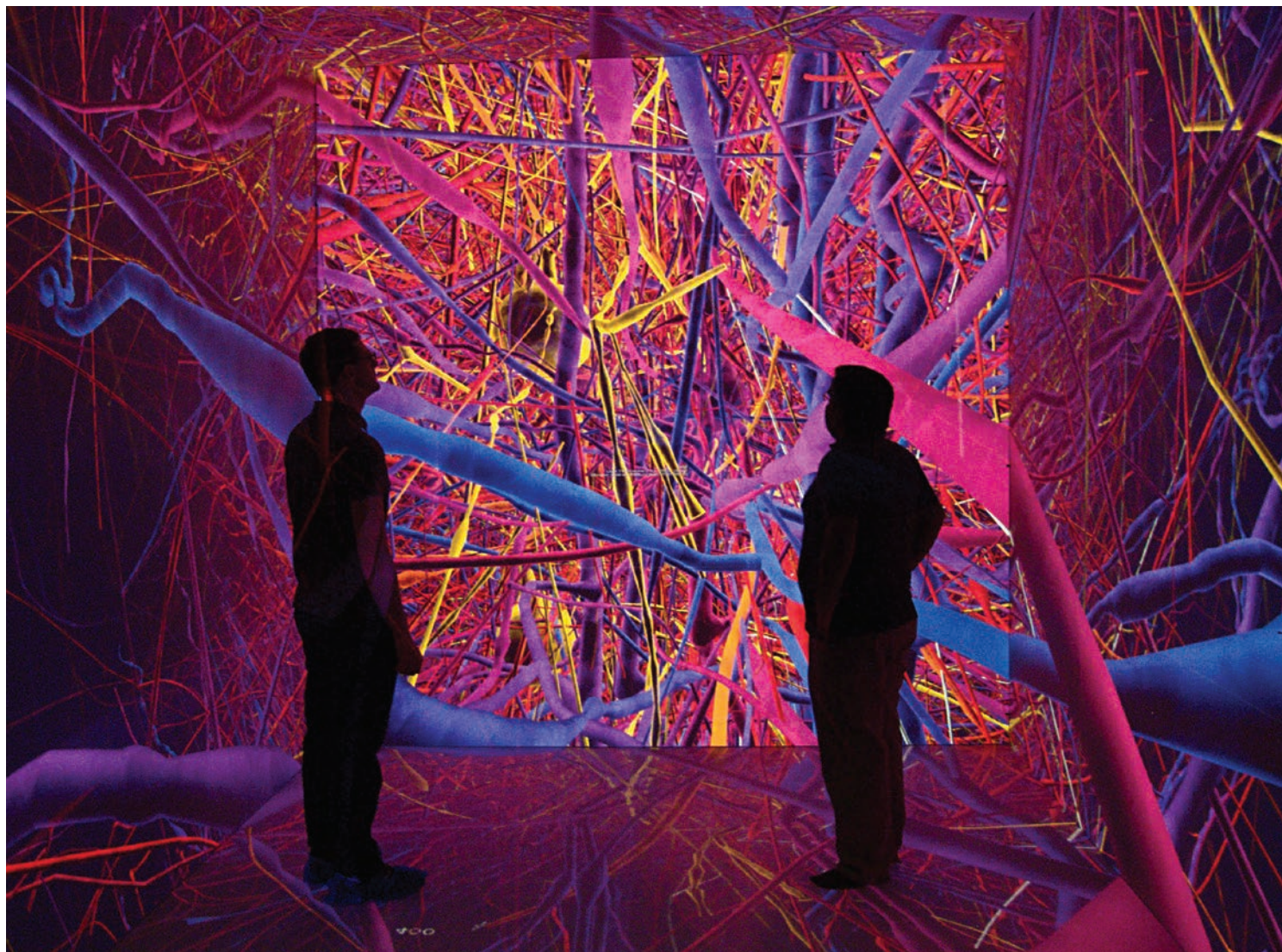
Looking at chemistry, the index shows that KAUST's collaboration score with UK institutions — which represent the third largest group after the US and France — increased from 2.5 in 2012 to 10.8 in 2015. In 2015, KAUST's top UK collaborator by far in chemistry was Imperial College London (ICL). In 2015 the data show that this partnership with ICL is KAUST's top partnership overall that year.

"Now we are an ongoing institution we are trying to make sure that our researchers can choose who they work with."

King Saud University (KSU) in the capital, Riyadh, the country's third-largest contributor to the index, is not an exception to the country's research trends. In 2015, however, it increased its collaboration with Russian institutions, making that country its collaboration partner of choice behind China and the US. Its top three collaborating institutions were Fudan University in China, the Russian Academy of Sciences (RAS) and Novosibirsk State University (NSU), both of which are in Russia.

Materials chemist, Ahmed Elzatahry, is prominent in KSU's collaboration with Fudan University. Elzatahry developed a relationship with Dongyuan Zhao, one of the world's top scientists in the field of mesoporous materials, in 2010 when Elzatahry was working in Egypt. When he moved to KSU in 2012, he took his research relationship with Zhao with him. This relationship has led to Elzatahry co-authoring several papers with Zhao, co-supervisions of PhD theses for KSU students, and an ongoing collaboration between KSU and Zhao that continues even though Elzatahry has recently moved on to Qatar University.

KSU clearly sees benefits to working with their counterparts on the international stage. Like their more prominent counterparts, most of the other 18 Saudi institutions whose international collaborations led to index publications have seen both their overall index output and their collaboration scores increase. The policy seems to have paid back for Saudi Arabia, which is likely to continue its pursuit of international partners as it works to boost its science output. ■



KAUST is starting to play a more domestic role by opening its excellent facilities to researchers from other Saudi Arabian research institutes.

MAKING THE MOST OF LOCAL EXPERTISE

Collaborating close to home means solving mutual problems and forging regional networks.

BY NADIA EL-AWADY

In 2011, a team of 17 researchers, based entirely in Saudi Arabia and Oman, published a paper in *Nature Genetics* identifying, for the first time, a single gene mutation that causes a rare form of the autoimmune disorder, systemic lupus erythematosus (SLE). That collaboration was led by geneticist Fowzan Alkuraya from King Faisal Specialist Hospital and Research Center (KFSH&RC) in the Saudi capital, Riyadh.

The centre followed this breakthrough in 2013, when Alkuraya's team proposed a role for an RNA helicase in the development of orofacioidigital syndrome. More recently, in early 2015 a team from KFSH&RC and King Saud University (KSU) identified a genetic

mutation linked to congenital cranial dysinnervation disorder. These studies represent a trend of national and regional collaboration at the institute which is rare in the Kingdom. While the index shows Saudi Arabia has strong research collaborations with the United States, Asia and Europe, national and collaborations with other countries in the Middle East appear sporadic.

"I don't see why local collaboration should be equated with low-quality output."

KFSH&RC and Alfaisal University, with which Alkuraya is also affiliated, are among the few Saudi institutions showing concerted and consistent national and regional

collaborations that result in studies published in high-quality journals tracked by the Nature Index. In 2015, only six of the 16 institutions that took part in domestic or regional collaborations produced significant output in the index, with KFSH&RC leading in terms of domestic collaborations. KSU, Saudi Arabia's second largest domestic collaborating institution in 2015, has the largest output from collaborations within the region of all Saudi institutions for that year.

Between 2012 and 2015, KFSH&RC, Alfaisal University and KSU — the country's top three for domestic collaborations for that period — derived 31%, 36% and 5% of their overall collaboration scores from domestic collaborations, compared to 0.5% for both Saudi Arabia's top

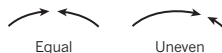
DOMESTIC RESEARCH NETWORK

Most of Saudi Arabia's collaborations take place outside the country, but domestic partnerships have been fostered by several of the Kingdom's top institutions.

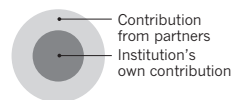
This diagram shows Saudi Arabia's domestic collaboration network in 2015. The circle sizes represent the CS each institution shares with other domestic partners.

LEGEND

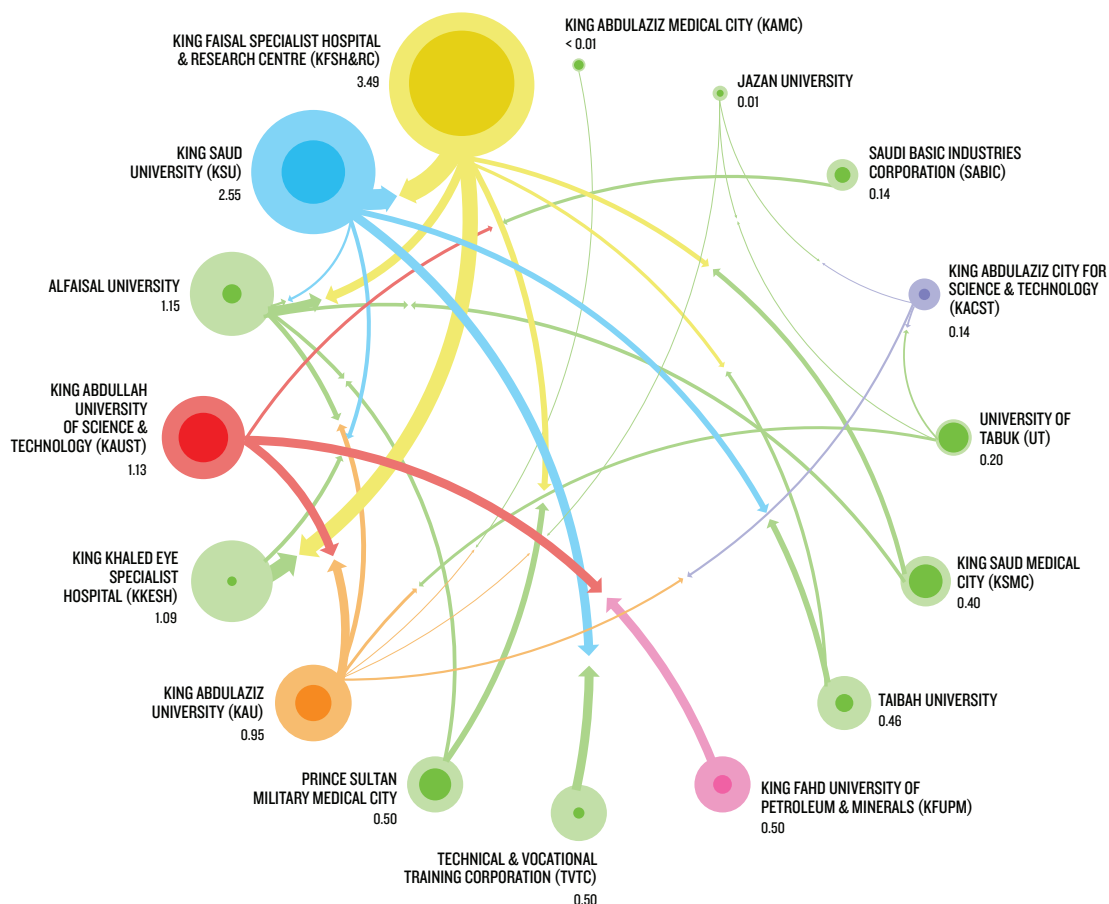
Contribution balance between institutions



Score split



CS is based on the sum of FC resulting from collaborative papers between any given institution and its domestic partners only.



international collaborators King Abdullah University for Science and Technology (KAUST) and King Abdulaziz University (KAU). "I try to focus on local and regional talent because I feel the responsibility of developing the research infrastructure as a good citizen of this region," says Alkuraya.

But Alkuraya isn't just being a good citizen. He truly believes in the wealth of local and regional talent. "I don't see why local and regional collaboration should be equated with low-quality research output," he says. "On the contrary, I think we have enough collaborative papers in high-impact journals to dispel such a myth."

In 2015, KFSH&RC researchers have had successful collaborations with 12 regional institutions, including, among others, Istanbul University in Turkey, Kuwait University, and six Saudi Arabian institutes, including KSU, King Khaled Eye Specialist Hospital and Alfaisal University. On the other hand, KAUST collaborated with only five regional and three Saudi Arabian institutes in the same year.

KAUST vice-president for research, Jean Fréchet, says his institution's dearth of local and regional partnerships was regrettable.

"Frankly, I think this was a mistake that we made. It is [just] that when we started KAUST, we were so busy getting the place started that we didn't have much time to look around," he said. "This is not sustainable and this is not something that we did on purpose."

TURNING EYES HOME

But this is changing, Fréchet says. In November 2015, KAUST invited researchers from King Fahd University of Petroleum and Minerals (KFUPM) to discuss potential joint research projects in the fields of cybersecurity, advanced computing and petroleum engineering. KAUST also has a large genomics programme with KFSH&RC, which is facilitated by King Abdulaziz City for Science and Technology (KACST).

KACST is also contributing funds to a KAUST-led project in solid state lighting, which includes collaboration with the University of California Santa Barbara, KFUPM and Effat University, a private women's university in Jeddah on the Kingdom's west coast. Five of KAUST's researchers also have small grants from Qatar Foundation to conduct research in collaboration with Qatar-based researchers.

Nevertheless, these domestic and regional

collaborations will probably continue to be small in comparison with KAUST's international work. "We have ambitious goals," says Fréchet. "We want to be rated as a top institution, and it's more tempting to rub elbows with the best."

This reasoning may be applicable for other Saudi Arabian institutes reaching out to collaborate with renowned institutions in the United States and Europe more often than their neighbours, where science output levels are more modest. Alkuraya believes there

"What kind of research infrastructure are you building by outsourcing the entire investigation?"

is a lack of incentive for local and regional researchers to collaborate. They often opt to publish more small papers in lesser-known journals because that helps earn promotion, rather than collaborating with other researchers and publishing in high-impact journals, which is why there are few papers for them appearing in the Nature Index. "There is also very little training that prepares investigators for team work," he adds.

KFSH&RC



KAU

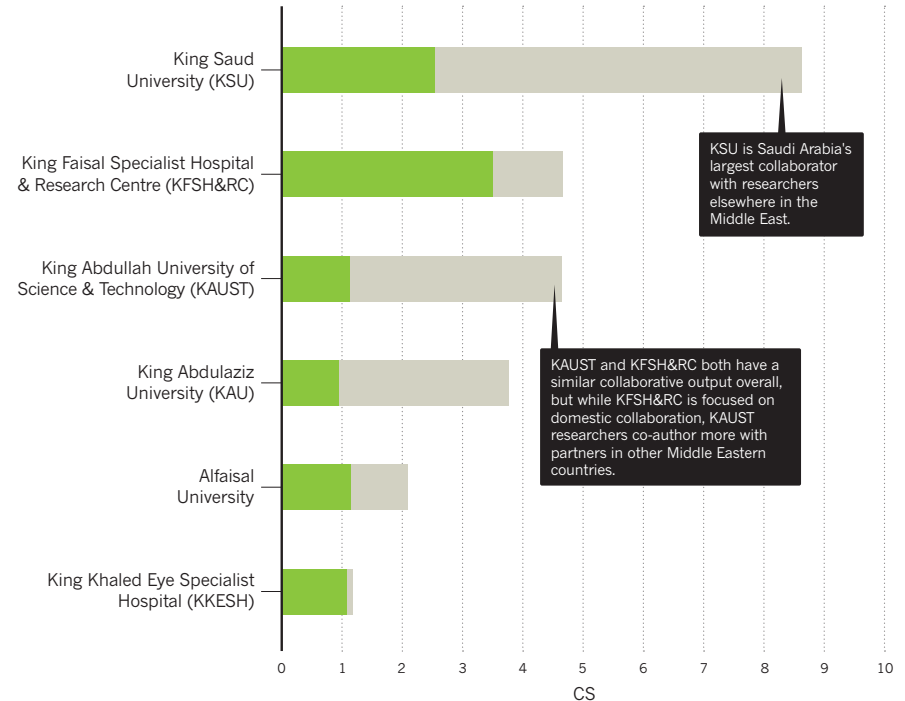


Genetic research is one of the disciplines driving domestic and regional research in Saudi Arabia.

REGIONAL COLLABORATIONS

In the wider Middle East region, these institutions stand above the rest in their collaborative output in the Nature Index in 2015. The bar chart below depicts each institution's CS for 2015 resulting from partnerships with counterparts in Saudi Arabia as well as in the rest of the Middle East.

● Saudi Arabia ● Middle East



Ahmed Elzatahry, an Egyptian material chemist who used to work at KSU, suggests that Saudi researchers have little to gain from regional collaborations. “They already have the people on site,” he says, referring to the large number of Arab expatriates from neighbouring countries living in Saudi Arabia. Elzatahry says Saudi institutions are often able to attract Arab researchers to come for the excellent facilities and generous research funding and salaries.

EQUAL COLLABORATIONS

KFSH&RC's Alkuraya seems to have found the secret to successful regional collaboration that leads to publication in high-quality journals.

“I'm incredibly lucky to have a very talented and dedicated group of individuals who work with me in the lab and a lot of credit really goes to them,” he says, emphasizing that his team members all have a strong sense of ownership of the research projects they are working on.

Alkuraya is scathing about what he calls a “pervasive trend” of shipping patient samples overseas to collaborators. “This trend may have worked to the advantage of individual investigators by getting their names well

published, but has been very damaging to the local research enterprise. What kind of research infrastructure are you building by outsourcing the entire scientific investigation?” he asks.

“To me, a true collaboration is made on equal footing where there is reciprocal and genuine exchange of expertise,” he says. “It gives me tremendous pleasure to choose an expert who agrees to perform an assay for which his or her lab has the expertise.”

Neurogeneticist Mustafa Salih was one of Alkuraya's former professors at KSU, and one of a few prominent Saudi researchers who have domestic and regional collaborations in the index. In 2015, KSU's domestic collaborations represented just 5% of its total collaborations. Salih believes studies in genetics are a main driver of Saudi Arabian local and regional collaborations. Between 2012 and 2015, life sciences collaborations in the index between Saudi institutes and their domestic and regional counterparts have indeed been the most productive.

Salih's domestic and regional collaborations include co-authored papers with researchers in Saudi Arabia, Turkey and Jordan related to retinal dystrophy and hereditary spastic paraplegia. “The collaborators in Turkey and

Jordan were interested in the same autosomal recessive neurogenetic disorders that have high prevalence in Saudi Arabia and the region,” he explains. Other top countries in the region with whom KSU-affiliated researchers have collaborated in the index between 2012 and 2015 include Egypt, Turkey and Cyprus.

KFSH&RC has a history of research successes, says Alkuraya. In the late 1990s, Mohamed Rashed was among those who pioneered the use of mass spectrometry to screen newborns for genetic metabolic disorders. Alkuraya says his team at KFSH&RC also published the first gene mapping study in the region that was performed solely by local talent. The study identified the mutation responsible for a rare disorder, Woodhouse–Sakati syndrome, first described among consanguineous Saudi Arabian families. His team was also the first to use locally generated exome sequencing to map a novel disease gene in the region.

Alkuraya points to these discoveries as good examples of local talent. “If we can't work together as a community of local and regional investigators, I don't know how we can sell ourselves as a bloc to the international research community,” he says. ■

A GUIDE TO THE NATURE INDEX

A description of the terminology and methodology used in this supplement, and a guide to the functionality available free online at natureindex.com.

The Nature Index is a database of author affiliations and institutional relationships. The index tracks contributions to articles published in a group of highly selective science journals, chosen by an independent group of active researchers.

The Nature Index provides absolute counts of publication productivity at the institutional and national level and, as such, is one indicator of global high-quality research output.

Data in the Nature Index are updated monthly, with the most recent 12 months of data made available under a Creative Commons licence at natureindex.com.

The database is compiled by Nature Publishing Group (NPG) in collaboration with Digital Science.

The list of journals tracked by the Nature Index is under review, and from 2016 will be extended to include the clinical sciences.

NATURE INDEX METRICS

There are four measures provided by the Nature Index to track affiliation data. The simplest is the **article count (AC)**. A country or institution is given an AC of 1 for each article that has at least one author from that country or institution. This is the case whether an article has one or a hundred authors, and it means that the same article can contribute to the AC of multiple countries or institutions.

To get a sense of a country or institution's contribution to an article, and to remove the possibility of counting articles more than once, the Nature Index uses the **fractional count (FC)**, which takes into account the relative contribution of each author to an article. The total FC available per paper is 1, which is shared between all authors under the assumption that each contributed equally. For instance, a paper with 10 authors means that each author receives an FC of 0.1. For authors who have joint affiliations, the individual FC is then split equally between each affiliation.

The third measure used is the weighted fractional count (WFC), which applies a weighting to the FC to adjust for the overrepresentation of papers in astronomy and astrophysics. The four journals in these disciplines publish about 50% of all papers in international journals in this field — approximately five times the equivalent percentage for other fields. Therefore, although the data for astronomy and astrophysics are compiled in the same way as for all other disciplines, articles from these journals are assigned one-fifth the weight of

natureindex.com users can search for specific institutions or countries and generate their own reports, ordered by article count (AC), fractional count (FC) or weighted fractional count (WFC).

Each query will return a profile page that lists the country or institution's recent research outputs, from which it is possible to drill down for more information. For example, articles can be displayed by journal, and then by article title. As in the supplement, research outputs are organized by subject area. The profile page also lists the institution or country's top collaborators, as well as its relationship with other research organizations.

other articles (i.e., the FC is multiplied by 0.2 to derive the WFC).

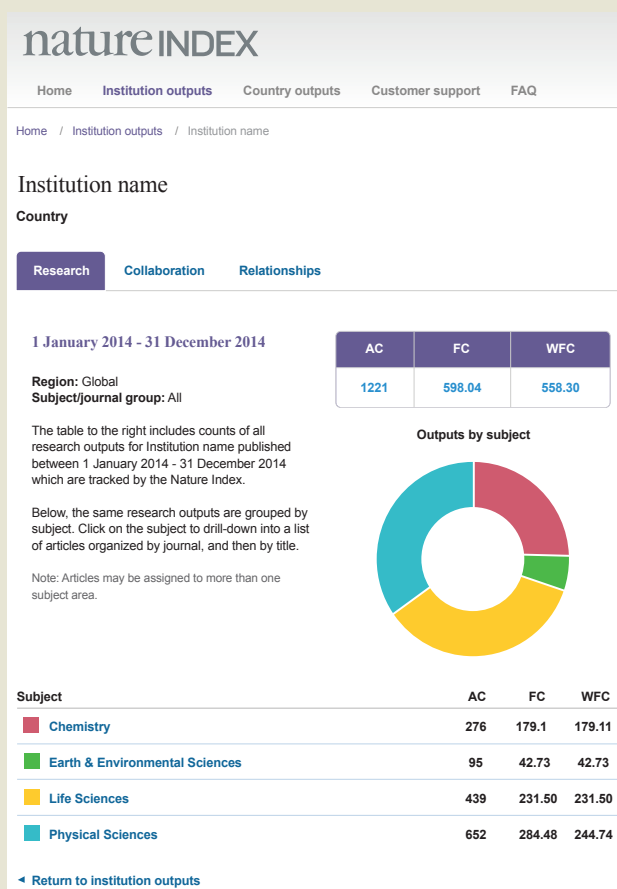
The total FC or WFC for an institution is calculated by summing the FC or WFC for individual authors. The fourth measure is the collaboration score (see The Supplement).

The process is similar for countries, although complicated by the fact that some institutions have overseas labs that will be counted towards their host country totals. What's more, there is great variability in the way authors present their affiliations. Every effort is made to count affiliations consistently, with a background of reasonable assumptions.

For more information on how the affiliation information is processed and counted, please see the FAQ section at natureindex.com.

NATUREINDEX.COM

A global indicator of high-quality research



THE SUPPLEMENT

Nature Index 2016 Saudi Arabia is based on data from the Nature Index, covering articles published during four consecutive years from 1 January 2012 to 31 December 2015.

Most analyses within the supplement use WFC as the primary metric, as it provides a more even basis for comparison across multiple disciplines, and in determining the relative contribution of each city or institution. Some sections and graphics also refer to collaboration score. This is a relatively new metric that is derived by adding the FC for all the bilateral relationships for that institution or country. If institution A has relationships with two others, B and C, then the collaboration score is the sum of FC for A + B and A + C. ■